1. Unit 2 Waves and Acoustics
    1. Chapter 15 Oscillations
        1. Introduction
        2. Simple Harmonic Motion
        3. Energy in Simple Harmonic Motion
        4. Comparing Simple Harmonic Motion and Circular Motion
        5. Pendulums
        6. Damped Oscillations
        7. Forced Oscillations
    2. Chapter 16 Waves
        1. Introduction
        2. Traveling Waves
        3. Mathematics of Waves
        4. Wave Speed on a Stretched String
        5. Energy and Power of a Wave
        6. Interference of Waves
        7. Standing Waves and Resonance
    3. Chapter 17 Sound
        1. Introduction
        2. Sound Waves
        3. Speed of Sound
        4. Sound Intensity
        5. Normal Modes of a Standing Sound Wave
        6. Sources of Musical Sound
        7. Beats
        8. The Doppler Effect
        9. Shock Waves
2. Unit 4 Electricity and Magnetism
    1. Chapter 22 Electric Charges and Fields
        1. Introduction

    2. Electric Charge
    3. Conductors, Insulators, and Charging by Induction
    4. Coulomb's Law
    5. Electric Field
    6. Calculating Electric Fields of Charge Distributions
    7. Electric Field Lines
    8. Electric Dipoles
2. Chapter 23 Gauss's Law
    1. Introduction
    2. Electric Flux
    3. Explaining Gauss’s Law
    4. Applying Gauss’s Law
    5. Conductors in Electrostatic Equilibrium
3. Chapter 24 Electric Potential
    1. Introduction
    2. Electric Potential Energy
    3. Electric Potential and Potential Difference
    4. Calculations of Electric Potential
    5. Determining Field from Potential
    6. Equipotential Surfaces and Conductors
    7. Applications of Electrostatics
4. Chapter 25 Capacitance
    1. Introduction
    2. Capacitors and Capacitance
    3. Capacitors in Series and in Parallel
    4. Energy Stored in a Capacitor
    5. Capacitor with a Dielectric
    6. Molecular Model of a Dielectric
5. Chapter 26 Current and Resistance
    1. Introduction
    2. Electrical Current
    3. Model of Conduction in Metals

    4. Resistivity and Resistance
    5. Ohm's Law
    6. Electrical Energy and Power
    7. Superconductors
6. Chapter 27 Direct-Current Circuits
    1. Introduction
    2. Electromotive Force
    3. Resistors in Series and Parallel
    4. Kirchhoff's Rules
    5. Electrical Measuring Instruments
    6. RC Circuits
    7. Household Wiring and Electrical Safety
7. Chapter 28 Magnetic Forces and Fields
    1. Introduction
    2. Magnetism and Its Historical Discoveries
    3. Magnetic Fields and Lines
    4. Motion of a Charged Particle in a Magnetic Field
    5. Magnetic Force on a Current-Carrying Conductor
    6. Force and Torque on a Current Loop
    7. The Hall Effect
    8. Applications of Magnetic Forces and Fields
8. Chapter 29 Sources of Magnetic Fields
    1. Introduction
    2. The Biot-Savart Law
    3. Magnetic Field Due to a Thin Straight Wire
    4. Magnetic Force between Two Parallel Currents
    5. Magnetic Field of a Current Loop
    6. Ampère’s Law
    7. Solenoids and Toroids
    8. Magnetism in Matter
9. Chapter 30 Electromagnetic Induction
    1. Introduction

        2. Faraday’s Law
        3. Lenz's Law
        4. Motional Emf
        5. Induced Electric Fields
        6. Eddy Currents
        7. Electric Generators and Back Emf
        8. Applications of Electromagnetic Induction
    10. Chapter 31 Inductance
        1. Introduction
        2. Mutual Inductance
        3. Self-Inductance and Inductors
        4. Energy in a Magnetic Field
        5. RL Circuits
        6. Oscillations in an LC Circuit
        7. RLC Series Circuits
    11. Chapter 32 Alternating-Current Circuits
        1. Introduction
        2. AC Sources
        3. Simple AC Circuits
        4. RLC Series Circuits with AC
        5. Power in an AC Circuit
        6. Resonance in an AC Circuit
        7. Transformers
    12. Chapter 33 Electromagnetic Waves
        1. Introduction
        2. Maxwell’s Equations and Electromagnetic Waves
        3. Plane Electromagnetic Waves
        4. Energy Carried by Electromagnetic Waves
        5. Momentum and Radiation Pressure
        6. The Electromagnetic Spectrum
3. Unit 5 Optics
    1. Chapter 34 The Nature of Light

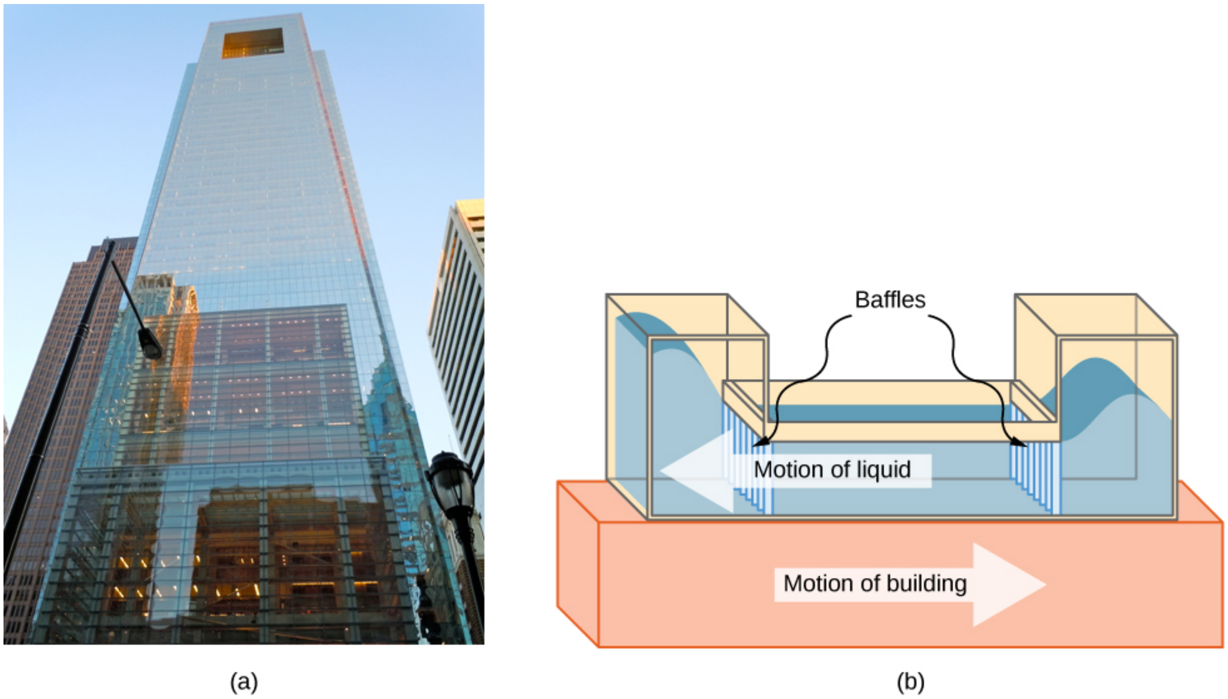
    1. Introduction
    2. The Propagation of Light
    3. The Law of Reflection
    4. Refraction
    5. Total Internal Reflection
    6. Dispersion
    7. Huygens’s Principle
    8. Polarization
2. Chapter 35 Geometric Optics and Image Formation
    1. Introduction
    2. Images Formed by Plane Mirrors
    3. Spherical Mirrors
    4. Images Formed by Refraction
    5. The Eye
    6. The Camera
    7. The Simple Magnifier
    8. Microscopes and Telescopes
3. Chapter 36 Interference
    1. Introduction
    2. Young's Double-Slit Interference
    3. Mathematics of Interference
    4. Multiple-Slit Interference
    5. Interference in Thin Films
    6. The Michelson Interferometer
4. Chapter 37 Diffraction
    1. Introduction
    2. Single-Slit Diffraction
    3. Intensity in Single-Slit Diffraction
    4. Double-Slit Diffraction
    5. Circular Apertures and Resolution
    6. X-Ray Diffraction
    7. Holography

4. Units
5. Conversion Factors
6. Fundamental Constants
7. Astronomical Data
8. Chemistry
9. The Greek Alphabet

Introduction class="introduction"

(a) The Comcast Building in Philadelphia, Pennsylvania, looming high above the skyline, is approximately 305 meters (1000 feet) tall. At this height, the top floors can oscillate back and forth due to seismic activity and fluctuating winds. (b) Shown above is a schematic drawing of a tuned, liquid-column mass damper, installed at the top of the Comcast, consisting of a 300,000-gallon reservoir of water to

reduce
oscillations.



(a) (b)

We begin the study of oscillations with simple systems of pendulums and springs. Although these systems may seem quite basic, the concepts involved have many real-life applications. For example, the Comcast Building in Philadelphia, Pennsylvania, stands approximately 305 meters (1000 feet) tall. As buildings are built taller, they can act as inverted, physical pendulums, with the top floors oscillating due to seismic activity and fluctuating winds. In the Comcast Building, a tuned-mass damper is used to reduce the oscillations. Installed at the top of the building is a tuned, liquid-column mass damper, consisting of a 300,000-gallon reservoir of water. This U-shaped tank allows the water to oscillate freely at a frequency that matches the natural frequency of the building. Damping is provided by tuning the turbulence levels in the moving water using baffles.

Simple Harmonic Motion

By the end of this section, you will be able to:

- Define the terms period and frequency
- List the characteristics of simple harmonic motion
- Explain the concept of phase shift
- Write the equations of motion for the system of a mass and spring undergoing simple harmonic motion
- Describe the motion of a mass oscillating on a vertical spring

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time ([link]). The string vibrates around an equilibrium position, and one oscillation is completed when the string starts from the initial position, travels to one of the extreme positions, then to the other extreme position, and returns to its initial position. We define **periodic motion** to be any motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by a child swinging on a swing. In this section, we study the basic characteristics of oscillations and their mathematical description.

When a guitar string is plucked, the string oscillates up and down in periodic motion. The vibrating string causes the surrounding air molecules to oscillate, producing sound waves. (credit: Yutaka Tsutano)

Period and Frequency in Oscillations

In the absence of friction, the time to complete one oscillation remains constant and is called the **period (*T*)**. Its units are usually seconds, but may be any convenient unit of time. The word ‘period’ refers to the time for some event whether repetitive or not, but in this chapter, we shall deal primarily in periodic motion, which is by definition repetitive.

A concept closely related to period is the frequency of an event. **Frequency (*f*)** is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

Note:
Equation:

$$f=\frac{1}{T}.$$

The SI unit for frequency is the *hertz* (Hz) and is defined as one *cycle per second*:
Equation:

$$1\text{ Hz}=1\frac{\text{cycle}}{\text{s}}\;\text{ or }\;1\text{ Hz}=\frac{1}{\text{s}}=1\text{ s}^{-1}.$$

A cycle is one complete **oscillation**.

Example:
Determining the Frequency of Medical Ultrasound
Ultrasound machines are used by medical professionals to make images for examining internal organs of the body. An ultrasound machine emits high-frequency sound waves, which reflect off the organs, and a computer receives the waves, using them to create a picture. We can use the formulas presented in this module to determine the frequency, based on what we know about oscillations. Consider a medical imaging device that produces ultrasound by oscillating with a period of $0.400\ \mu\text{s}$. What is the frequency of this oscillation?
Strategy
The period (*T*) is given and we are asked to find frequency (*f*).
Solution

Substitute $0.400\ \mu\text{s}$ for T in $f = \frac{1}{T}$:

Equation:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6}\ \text{s}}.$$

Solve to find

Equation:

$$f = 2.50 \times 10^{6}\ \text{Hz}.$$

Significance

This frequency of sound is much higher than the highest frequency that humans can hear (the range of human hearing is 20 Hz to 20,000 Hz); therefore, it is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.

Characteristics of Simple Harmonic Motion

A very common type of periodic motion is called **simple harmonic motion (SHM)**. A system that oscillates with SHM is called a **simple harmonic oscillator**.

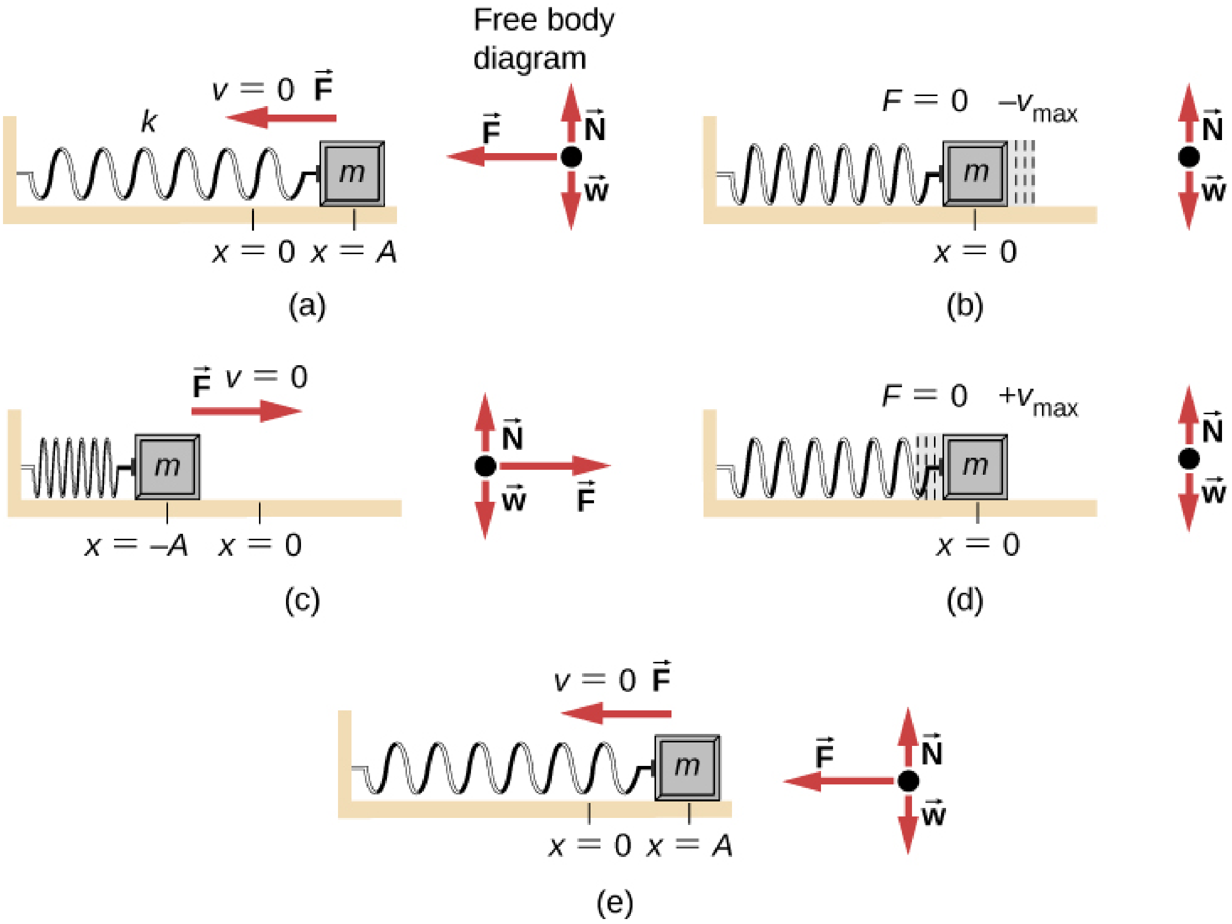
Note:

Simple Harmonic Motion

In simple harmonic motion, the acceleration of the system, and therefore the net force, is proportional to the displacement and acts in the opposite direction of the displacement.

A good example of SHM is an object with mass *m* attached to a spring on a frictionless surface, as shown in [link]. The object oscillates around the equilibrium position, and the net force on the object is equal to the force provided by the spring. This force obeys Hooke’s law $F_s = -kx$, as discussed in a previous chapter.

If the net force can be described by Hooke’s law and there is no *damping* (slowing down due to friction or other nonconservative forces), then a simple harmonic oscillator oscillates with equal displacement on either side of the equilibrium position, as shown for an object on a spring in [link]. The maximum displacement from equilibrium is called the **amplitude (A)**. The units for amplitude and displacement are the same but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters.


Free body diagram
$v = 0$
$\vec{F}$
k
m
$\vec{F}$
$\vec{N}$
$\vec{w}$
$x = 0$
$x = A$
(a)
$F = 0$
$-v_{max}$
m
$\vec{N}$
$\vec{w}$
$x = 0$
(b)
$\vec{F}$
$v = 0$
m
$x = -A$
$x = 0$
$\vec{N}$
$\vec{w}$
$\vec{F}$
(c)
$F = 0$
$+v_{max}$
m
$\vec{N}$
$\vec{w}$
$x = 0$
(d)
$v = 0$
$\vec{F}$
m
$x = 0$
$x = A$
$\vec{F}$
$\vec{N}$
$\vec{w}$
(e)

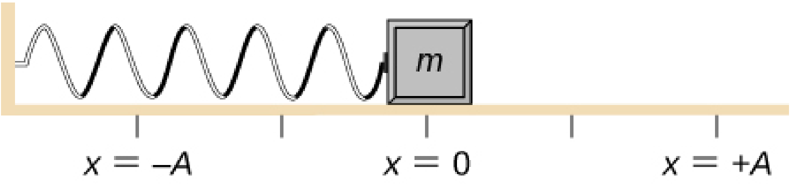
An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. In the above set of figures, a mass is attached to a spring and placed on a frictionless table. The other end of the spring is attached to the wall. The position of the mass, when the spring is neither stretched nor compressed, is marked as $x = 0$ and is the equilibrium position. (a) The mass is displaced to a position $x = A$ and released from rest. (b) The mass accelerates as it moves in the negative *x*-direction, reaching a maximum negative velocity at $x = 0$. (c) The mass continues to move in the negative *x*-direction, slowing until it comes to a stop at $x = -A$. (d) The mass now begins to accelerate in the positive *x*-direction, reaching a positive maximum velocity at $x = 0$. (e) The mass then continues to move in the positive direction until it stops at $x = A$. The mass continues in SHM that has an amplitude *A* and a period *T*. The object's maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

What is so significant about SHM? For one thing, the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, oscillates with the same frequency whether plucked gently or hard.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large **force constant (k)**, which causes the system to have a smaller period. For example, you can adjust a diving board's stiffness—the stiffer it is, the faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one. In fact, the mass m and the force constant k are the *only* factors that affect the period and frequency of SHM. To derive an equation for the period and the frequency, we must first define and analyze the equations of motion. Note that the force constant is sometimes referred to as the *spring constant*.

Equations of SHM

Consider a block attached to a spring on a frictionless table ([link]). The **equilibrium position** (the position where the spring is neither stretched nor compressed) is marked as $x = 0$. At the equilibrium position, the net force is zero.



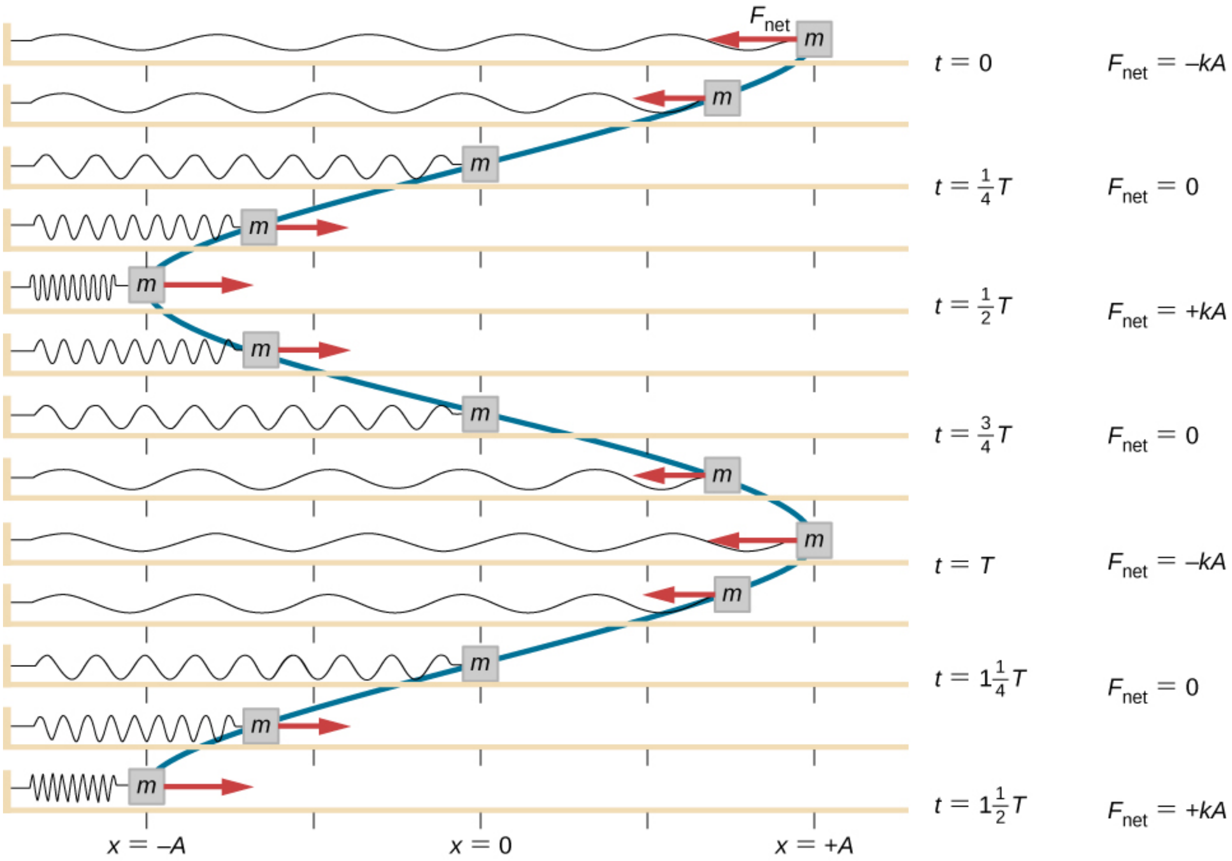
A block is attached to a spring and placed on a frictionless table. The equilibrium position, where the spring is neither extended nor compressed, is marked as $x = 0$.

Work is done on the block to pull it out to a position of $x = +A$, and it is then released from rest. The maximum *x*-position (*A*) is called the amplitude of the motion. The block begins to oscillate in SHM between $x = +A$ and $x = -A$, where *A* is the amplitude of the motion and *T* is the period of the oscillation. The period is the time for one oscillation. [link] shows the motion of the block as it completes one and a half oscillations after release. [link] shows a plot of the position of the block versus time. When the position is plotted versus time, it is clear that the data can be modeled by a cosine function with an amplitude *A* and a period *T*. The cosine function cosθ repeats every multiple of 2π, whereas the motion of the block repeats every period *T*. However, the function $\cos\left(\frac{2\pi}{T}t\right)$ repeats every integer multiple of the period. The maximum of the cosine function is one, so it is necessary to multiply the cosine function by the amplitude *A*.

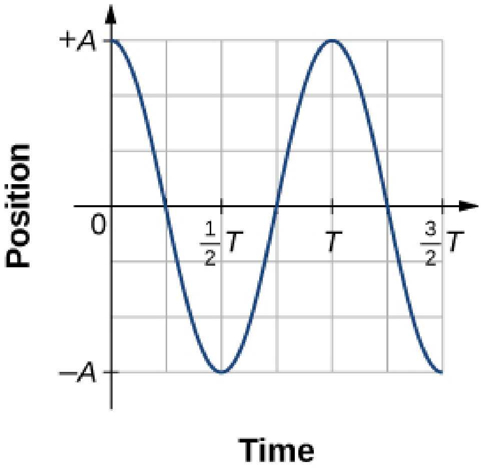
Note:
Equation:

$$x\left(t\right) = A\cos\left(\frac{2\pi}{T}t\right) = A\cos\left(\omega t\right).$$

Recall from the chapter on rotation that the angular frequency equals $\omega = \frac{d\theta}{dt}$. In this case, the period is constant, so the angular frequency is defined as 2π divided by the period, $\omega = \frac{2\pi}{T}$.

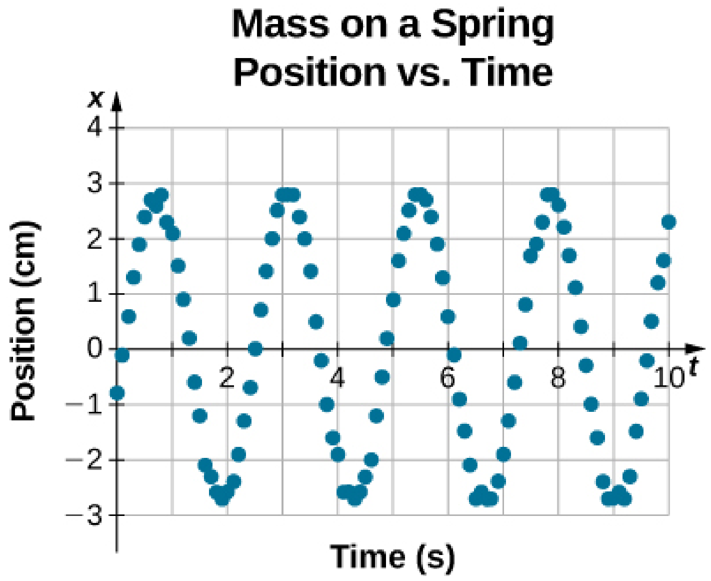

F_{net}
m
$t = 0$
$F_{net} = -kA$
$t = \frac{1}{4}T$
$F_{net} = 0$
$t = \frac{1}{2}T$
$F_{net} = +kA$
$t = \frac{3}{4}T$
$F_{net} = 0$
$t = T$
$F_{net} = -kA$
$t = 1\frac{1}{4}T$
$F_{net} = 0$
$t = 1\frac{1}{2}T$
$F_{net} = +kA$
$x = -A$
$x = 0$
$x = +A$

A block is attached to one end of a spring and placed on a frictionless table. The other end of the spring is anchored to the wall. The equilibrium position, where the net force equals zero, is marked as $x = 0\,\mathrm{m}$. Work is done on the block, pulling it out to $x = +A$, and the block is released from rest. The block oscillates between $x = +A$ and $x = -A$. The force is also shown as a vector.



A graph of the position of the block shown in [link] as a function of time. The position can be modeled as a periodic function, such as a cosine or sine function.

The equation for the position as a function of time $x\left(t\right)=A\cos\left(\omega t\right)$ is good for modeling data, where the position of the block at the initial time $t=0.00\,\text{s}$ is at the amplitude A and the initial velocity is zero. Often when taking experimental data, the position of the mass at the initial time $t=0.00\,\text{s}$ is not equal to the amplitude and the initial velocity is not zero. Consider 10 seconds of data collected by a student in lab, shown in [link].



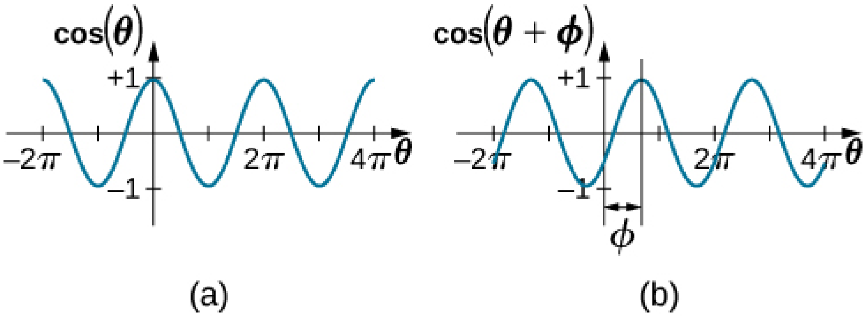
Data collected by a student in lab indicate the position of a block attached to a spring, measured with a sonic range finder. The data are collected starting at time $t = 0.00\text{s}$, but the initial position is near position $x \approx -0.80 \text{ cm} \neq 3.00 \text{ cm}$, so the initial position does not equal the amplitude $x_0 = +A$. The velocity is the time derivative of the position, which is the slope at a point on the graph of position versus time. The velocity is not $v = 0.00 \text{ m/s}$ at time $t = 0.00 \text{ s}$, as evident by the slope of the graph of position versus time, which is not zero at the initial time.

The data in [link] can still be modeled with a periodic function, like a cosine function, but the function is shifted to the right. This shift is known as a **phase shift** and is usually represented by the Greek letter phi (ϕ). The equation of the position as a function of time for a block on a spring becomes

Equation:

$$x\left(t\right) = A\cos\left(\omega t + \phi\right).$$

This is the generalized equation for SHM where *t* is the time measured in seconds, ω is the angular frequency with units of inverse seconds, *A* is the amplitude measured in meters or centimeters, and ϕ is the phase shift measured in radians ([link]). It should be noted that because sine and cosine functions differ only by a phase shift, this motion could be modeled using either the cosine or sine function.



(a) A cosine function. (b) A cosine function shifted to the right by an angle ϕ. The angle ϕ is known as the phase shift of the function.

The velocity of the mass on a spring, oscillating in SHM, can be found by taking the derivative of the position equation:

Equation:

$$v\left(t\right)=\frac{dx}{dt}=\frac{d}{dt}\left(A\cos\left(\omega t+\phi\right)\right)=-A\omega\sin\left(\omega t+\varphi\right)=-v_{\max}\sin\left(\omega t+\phi\right).$$

Because the sine function oscillates between –1 and +1, the maximum velocity is the amplitude times the angular frequency, $v_{\max}=A\omega$. The maximum velocity occurs at the equilibrium position $\left(x=0\right)$ when the mass is moving toward $x=+A$. The maximum velocity in the negative direction is attained at the equilibrium position $\left(x=0\right)$ when the mass is moving toward $x=-A$ and is equal to $-v_{\max}$.

The acceleration of the mass on the spring can be found by taking the time derivative of the velocity:

Equation:

$$a\left(t\right)=\frac{dv}{dt}=\frac{d}{dt}\left(-A\omega\sin\left(\omega t+\phi\right)\right)=-A\omega^2\cos\left(\omega t+\varphi\right)=-a_{\max}\cos\left(\omega t+\phi\right).$$

The maximum acceleration is $a_{\max} = A\omega^2$. The maximum acceleration occurs at the position$(x = -A)$, and the acceleration at the position $(x = -A)$ and is equal to $-a_{\max}$.

Summary of Equations of Motion for SHM

In summary, the oscillatory motion of a block on a spring can be modeled with the following equations of motion:

Note:
Equation:

$$x\left(t\right) = A\cos\left(\omega t + \phi\right)$$

Equation:

$$v\left(t\right) = -v_{\max}\sin\left(\omega t + \phi\right)$$

Equation:

$$a\left(t\right) = -a_{\max}\cos\left(\omega t + \phi\right)$$

Equation:

$$x_{\max} = A$$

Equation:

$$v_{\max} = A\omega$$

Equation:

$$a_{\max} = A\omega^2.$$

Here, A is the amplitude of the motion, T is the period, ϕ is the phase shift, and $\omega = \frac{2\pi}{T} = 2\pi f$ is the angular frequency of the motion of the block.

Example:

Determining the Equations of Motion for a Block and a Spring

A 2.00-kg block is placed on a frictionless surface. A spring with a force constant of $k = 32.00\,\text{N/m}$ is attached to the block, and the opposite end of the spring is attached to the wall. The spring can be compressed or extended. The equilibrium position is marked as $x = 0.00\,\text{m}$.

Work is done on the block, pulling it out to $x = +0.02\,\text{m}$. The block is released from rest and oscillates between $x = +0.02\,\text{m}$ and $x = -0.02\,\text{m}$. The period of the motion is 1.57 s. Determine the equations of motion.

Strategy

We first find the angular frequency. The phase shift is zero, $\phi = 0.00\,\text{rad}$, because the block is released from rest at $x = A = +0.02\,\text{m}$. Once the angular frequency is found, we can determine the maximum velocity and maximum acceleration.

Solution

The angular frequency can be found and used to find the maximum velocity and maximum acceleration:

Equation:

$$\begin{aligned} \omega &= \frac{2\pi}{1.57\,\text{s}} = 4.00\,\text{s}^{-1}; \\ v_{\text{max}} &= A\omega = 0.02\text{m}\left(4.00\,\text{s}^{-1}\right) = 0.08\,\text{m/s}; \\ a_{\text{max}} &= A\omega^2 = 0.02\,\text{m}\left(4.00\,\text{s}^{-1}\right)^2 = 0.32\,\text{m/s}^2. \end{aligned}$$

All that is left is to fill in the equations of motion:

Equation:

$$\begin{aligned} x\left(t\right) &= A\cos\left(\omega t + \phi\right) = (0.02\,\text{m})\cos\left(4.00\,\text{s}^{-1}t\right); \\ v\left(t\right) &= -v_{\text{max}}\sin\left(\omega t + \phi\right) = (-0.08\,\text{m/s})\sin\left(4.00\,\text{s}^{-1}t\right); \\ a\left(t\right) &= -a_{\text{max}}\cos\left(\omega t + \phi\right) = \left(-0.32\,\text{m/s}^2\right)\cos\left(4.00\,\text{s}^{-1}t\right). \end{aligned}$$

Significance

The position, velocity, and acceleration can be found for any time. It is important to remember that when using these equations, your calculator must be in radians mode.

The Period and Frequency of a Mass on a Spring

One interesting characteristic of the SHM of an object attached to a spring is that the angular frequency, and therefore the period and frequency of the motion, depend on only the mass and the force constant, and not on other factors such as the amplitude of the motion. We can use the equations of motion and Newton’s second law ($\vec{\mathbf{F}}_{\text{net}} = m\vec{\mathbf{a}}$) to find equations for the angular frequency, frequency, and period.

Consider the block on a spring on a frictionless surface. There are three forces on the mass: the weight, the normal force, and the force due to the spring. The only two forces that act perpendicular to the surface are the weight and the normal force, which have equal magnitudes and opposite directions, and thus sum to zero. The only force that acts parallel to the surface is the force due to the spring, so the net force must be equal to the force of the spring:

Equation:

$$\begin{array}{rcl}F_x & = & -kx; \\ ma & = & -kx; \\ m\frac{d^2x}{dt^2} & = & -kx; \\ \frac{d^2x}{dt^2} & = & -\frac{k}{m}x.\end{array}$$

Substituting the equations of motion for *x* and *a* gives us

Equation:

$$-A\omega^2 \cos(\omega t + \phi) = -\frac{k}{m} A \cos(\omega t + \phi).$$

Cancelling out like terms and solving for the angular frequency yields

Note:

Equation:

$$\omega = \sqrt{\frac{k}{m}}.$$

The angular frequency depends only on the force constant and the mass, and not the amplitude. The angular frequency is defined as $\omega = 2\pi/T$, which yields an equation for the period of the motion:

Note:
Equation:

$$T = 2\pi\sqrt{\frac{m}{k}}.$$

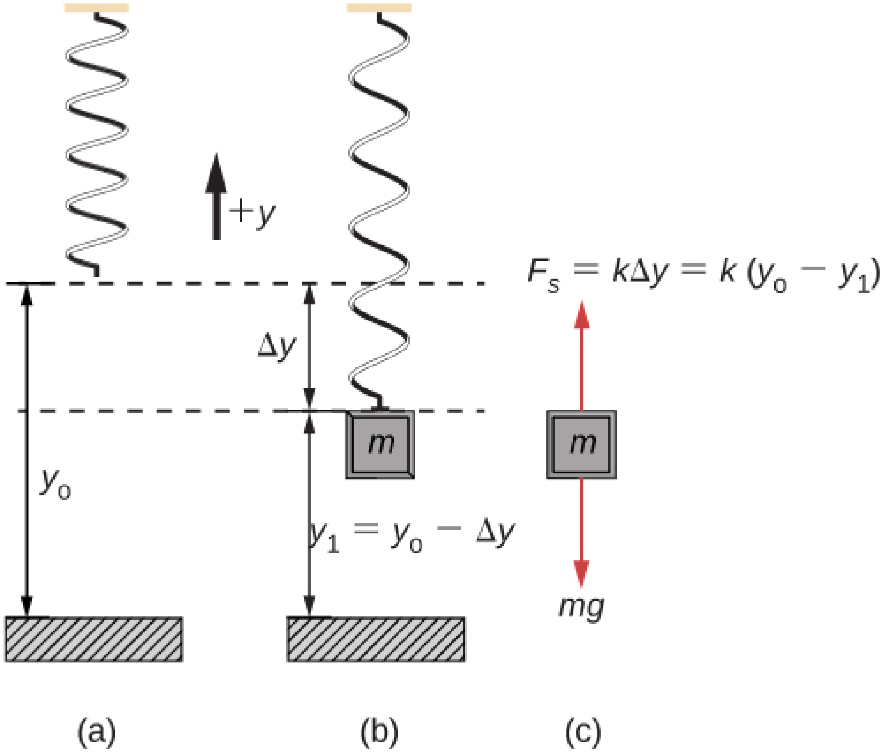
The period also depends only on the mass and the force constant. The greater the mass, the longer the period. The stiffer the spring, the shorter the period. The frequency is

Note:
Equation:

$$f = \frac{1}{T} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}.$$

Vertical Motion and a Horizontal Spring

When a spring is hung vertically and a block is attached and set in motion, the block oscillates in SHM. In this case, there is no normal force, and the net effect of the force of gravity is to change the equilibrium position. Consider [link]. Two forces act on the block: the weight and the force of the spring. The weight is constant and the force of the spring changes as the length of the spring changes.



A spring is hung from the ceiling. When a block is attached, the block is at the equilibrium position where the weight of the block is equal to the force of the spring. (a) The spring is hung from the ceiling and the equilibrium position is marked as y_o. (b) A mass is attached to the spring and a new equilibrium position is reached ($y_1 = y_o - \Delta y$) when the force provided by the spring equals the weight of the mass. (c) The free-body diagram of the mass shows the two forces acting on the mass: the weight and the force of the spring.

When the block reaches the equilibrium position, as seen in [link], the force of the spring equals the weight of the block, $F_{\text{net}} = F_{\text{s}} - mg = 0$, where

Equation:

$$-k\left(-\Delta y\right) = mg.$$

From the figure, the change in the position is $\Delta y = y_0 - y_1$ and since $-k\left(-\Delta y\right) = mg$, we have

Equation:

$$k\left(y_0 - y_1\right) - mg = 0.$$

If the block is displaced and released, it will oscillate around the new equilibrium position. As shown in [link], if the position of the block is recorded as a function of time, the recording is a periodic function.

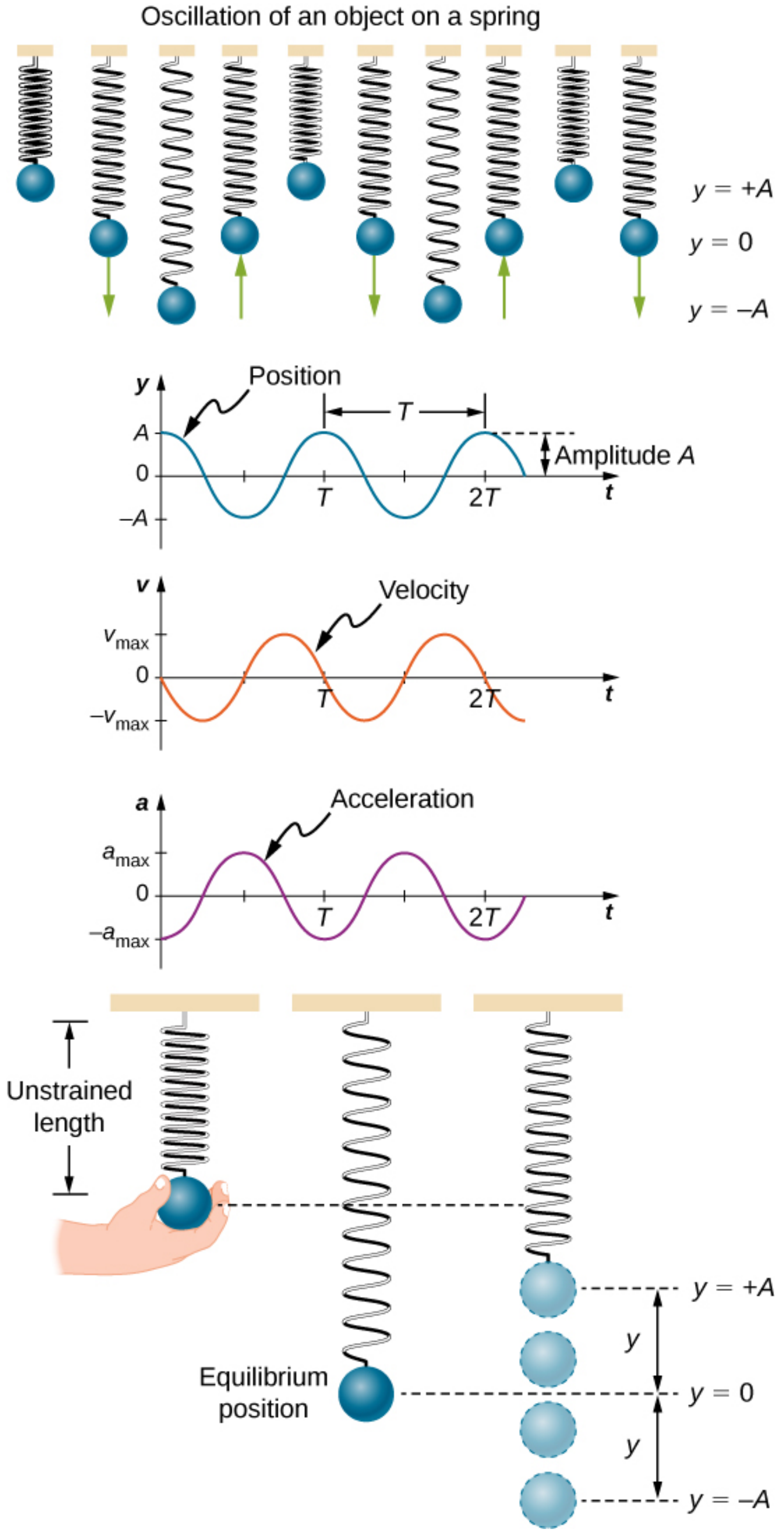
If the block is displaced to a position *y*, the net force becomes $F_{\text{net}} = k\left(y - y_0\right) - mg = 0$. But we found that at the equilibrium position, $mg = k\Delta y = ky_0 - ky_1$. Substituting for the weight in the equation yields
Equation:

$$F_{\text{net}} = ky - ky_0 - \left(ky_0 - ky_1\right) = -k\left(y - y_1\right).$$

Recall that y_1 is just the equilibrium position and any position can be set to be the point $y = 0.00\text{m}$. So let’s set y_1 to $y = 0.00\text{ m}$. The net force then becomes
Equation:

$$\begin{aligned} F_{\text{net}} &= -ky; \\ m\frac{d^2y}{dt^2} &= -ky. \end{aligned}$$

This is just what we found previously for a horizontally sliding mass on a spring. The constant force of gravity only served to shift the equilibrium location of the mass. Therefore, the solution should be the same form as for a block on a horizontal spring, $y\left(t\right) = A\cos\left(\omega t + \phi\right)$. The equations for the velocity and the acceleration also have the same form as for the horizontal case. Note that the inclusion of the phase shift means that the motion can actually be modeled using either a cosine or a sine function, since these two functions only differ by a phase shift.


Oscillation of an object on a spring
y = +A
y = 0
y = –A
Position
y
A
0
–A
T
T
2T
t
Amplitude A
Velocity
v
v max
0
–v max
T
2T
t
Acceleration
a
a max
0
–a max
T
2T
t
Unstrained length
Equilibrium position
y = +A
y
y = 0
y
y = –A

Graphs of $y(t)$, $v(t)$, and $a(t)$ versus t for the motion of an object on a vertical spring. The net force on the object can be described by Hooke’s law, so the object undergoes SHM. Note that the initial position has the vertical displacement at its maximum value *A*; *v* is initially zero and then negative as the object moves down; the initial acceleration is negative, back toward the equilibrium position and becomes zero at that point.

Summary

- Periodic motion is a repeating oscillation. The time for one oscillation is the period *T* and the number of oscillations per unit time is the frequency *f*. These quantities are related by $f=\frac{1}{T}$.
- Simple harmonic motion (SHM) is oscillatory motion for a system where the restoring force is proportional to the displacement and acts in the direction opposite to the displacement.
- Maximum displacement is the amplitude *A*. The angular frequency *ω*, period *T*, and frequency *f* of a simple harmonic oscillator are given by $\omega=\sqrt{\frac{k}{m}}$, $T=2\pi\sqrt{\frac{m}{k}},\text{ and }f=\frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where *m* is the mass of the system and *k* is the force constant.
- Displacement as a function of time in SHM is given by $x(t)=A\cos\left(\frac{2\pi}{T}t+\phi\right)=A\cos\left(\omega t+\phi\right)$.
- The velocity is given by $v(t)=-A\omega\sin\left(\omega t+\phi\right)=-v_{\text{max}}\sin\left(\omega t+\phi\right),\text{ where }\mathrm{v}_{\text{max}}=A\omega=A\sqrt{\frac{k}{m}}$.
- The acceleration is $a(t)=-A\omega^2\cos\left(\omega t+\phi\right)=-a_{\text{max}}\cos\left(\omega t+\phi\right)$, where $a_{\text{max}}=A\omega^2=A\frac{k}{m}$.

Conceptual Questions

Exercise:

Problem: What conditions must be met to produce SHM?

Solution:

The restoring force must be proportional to the displacement and act opposite to the direction of motion with no drag forces or friction. The frequency of oscillation does not depend on the amplitude.

Exercise:

Problem:

(a) If frequency is not constant for some oscillation, can the oscillation be SHM? (b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?

Exercise:

Problem:

Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

Solution:

Examples: Mass attached to a spring on a frictionless table, a mass hanging from a string, a simple pendulum with a small amplitude of motion. All of these examples have frequencies of oscillation that are independent of amplitude.

Exercise:

Problem:

Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a more pliable material.

Exercise:

Problem:

As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.

Solution:

Since the frequency is proportional to the square root of the force constant and inversely proportional to the square root of the mass, it is likely that the truck is heavily loaded, since the force constant would be the same whether the truck is empty or heavily loaded.

Exercise:

Problem:

Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

Problems

Exercise:

Problem:

Prove that using $x\left(t\right)=A\text{sin}\left(\omega t+\phi \right)$ will produce the same results for the period for the oscillations of a mass and a spring. Why do you think the cosine function was chosen?

Solution:

Proof

Exercise:

Problem: What is the period of 60.0 Hz of electrical power?

Exercise:

Problem:

If your heart rate is 150 beats per minute during strenuous exercise, what is the time per beat in units of seconds?

Solution:

0.400 s/beat

Exercise:

Problem:

Find the frequency of a tuning fork that takes $2.50\phantom{\rule{0.2em}{0ex}}\times \phantom{\rule{0.2em}{0ex}}{10}^{-3}$s to complete one oscillation.

Exercise:

Problem:

A stroboscope is set to flash every 8.00×10^{-5}s. What is the frequency of the flashes?

Solution:

12,500 Hz

Exercise:

Problem:

A tire has a tread pattern with a crevice every 2.00 cm. Each crevice makes a single vibration as the tire moves. What is the frequency of these vibrations if the car moves at 30.0 m/s?

Exercise:

Problem:

Each piston of an engine makes a sharp sound every other revolution of the engine. (a) How fast is a race car going if its eight-cylinder engine emits a sound of frequency 750 Hz, given that the engine makes 2000 revolutions per kilometer? (b) At how many revolutions per minute is the engine rotating?

Solution:

a. 340 km/hr; b. 11.3×10^3 rev/min

Exercise:

Problem:

A type of cuckoo clock keeps time by having a mass bouncing on a spring, usually something cute like a cherub in a chair. What force constant is needed to produce a period of 0.500 s for a 0.0150-kg mass?

Exercise:

Problem:

A mass m_0 is attached to a spring and hung vertically. The mass is raised a short distance in the vertical direction and released. The mass oscillates with a frequency f_0. If the mass is replaced with a mass nine times as large, and the experiment was repeated, what would be the frequency of the oscillations in terms of f_0 ?

Solution:

$f = \frac{1}{3} f_0$

Exercise:

Problem:

A 0.500-kg mass suspended from a spring oscillates with a period of 1.50 s. How much mass must be added to the object to change the period to 2.00 s?

Exercise:

Problem:

By how much leeway (both percentage and mass) would you have in the selection of the mass of the object in the previous problem if you did not wish the new period to be greater than 2.01 s or less than 1.99 s?

Solution:

0.009 kg; 2%

Glossary

amplitude (*A*)	maximum displacement from the equilibrium position of an object oscillating around the equilibrium position
equilibrium position	position where the spring is neither stretched nor compressed
force constant (*k*)	characteristic of a spring which is defined as the ratio of the force applied to the spring to the displacement caused by the force
frequency (*f*)	number of events per unit of time
oscillation	single fluctuation of a quantity, or repeated and regular fluctuations of a quantity, between two extreme values around an equilibrium or average value
periodic motion	

motion that repeats itself at regular time intervals

period (*T*)
time taken to complete one oscillation

phase shift
angle, in radians, that is used in a cosine or sine function to shift the function left or right, used to match up the function with the initial conditions of data

simple harmonic motion (SHM)
oscillatory motion in a system where the restoring force is proportional to the displacement, which acts in the direction opposite to the displacement

simple harmonic oscillator
a device that oscillates in SHM where the restoring force is proportional to the displacement and acts in the direction opposite to the displacement

Energy in Simple Harmonic Motion
By the end of this section, you will be able to:

- Describe the energy conservation of the system of a mass and a spring
- Explain the concepts of stable and unstable equilibrium points

To produce a deformation in an object, we must do work. That is, whether you pluck a guitar string or compress a car’s shock absorber, a force must be exerted through a distance. If the only result is deformation, and no work goes into thermal, sound, or kinetic energy, then all the work is initially stored in the deformed object as some form of potential energy.

Consider the example of a block attached to a spring on a frictionless table, oscillating in SHM. The force of the spring is a conservative force (which you studied in the chapter on potential energy and conservation of energy), and we can define a potential energy for it. This potential energy is the energy stored in the spring when the spring is extended or compressed. In this case, the block oscillates in one dimension with the force of the spring acting parallel to the motion:
Equation:

$$W=\int_{x_i}^{x_f} F_x dx=\int_{x_i}^{x_f} -kx dx=\left[-\frac{1}{2}kx^2\right]_{x_i}^{x_f}=-\left[\frac{1}{2}kx_f^2-\frac{1}{2}kx_i^2\right]=-\left[U_f-U_i\right]=-\Delta U.$$

When considering the energy stored in a spring, the equilibrium position, marked as $x_i = 0.00\,\mathrm{m}$, is the position at which the energy stored in the spring is equal to zero. When the spring is stretched or compressed a distance *x*, the potential energy stored in the spring is
Equation:

$$U=\frac{1}{2}kx^2.$$

Energy and the Simple Harmonic Oscillator

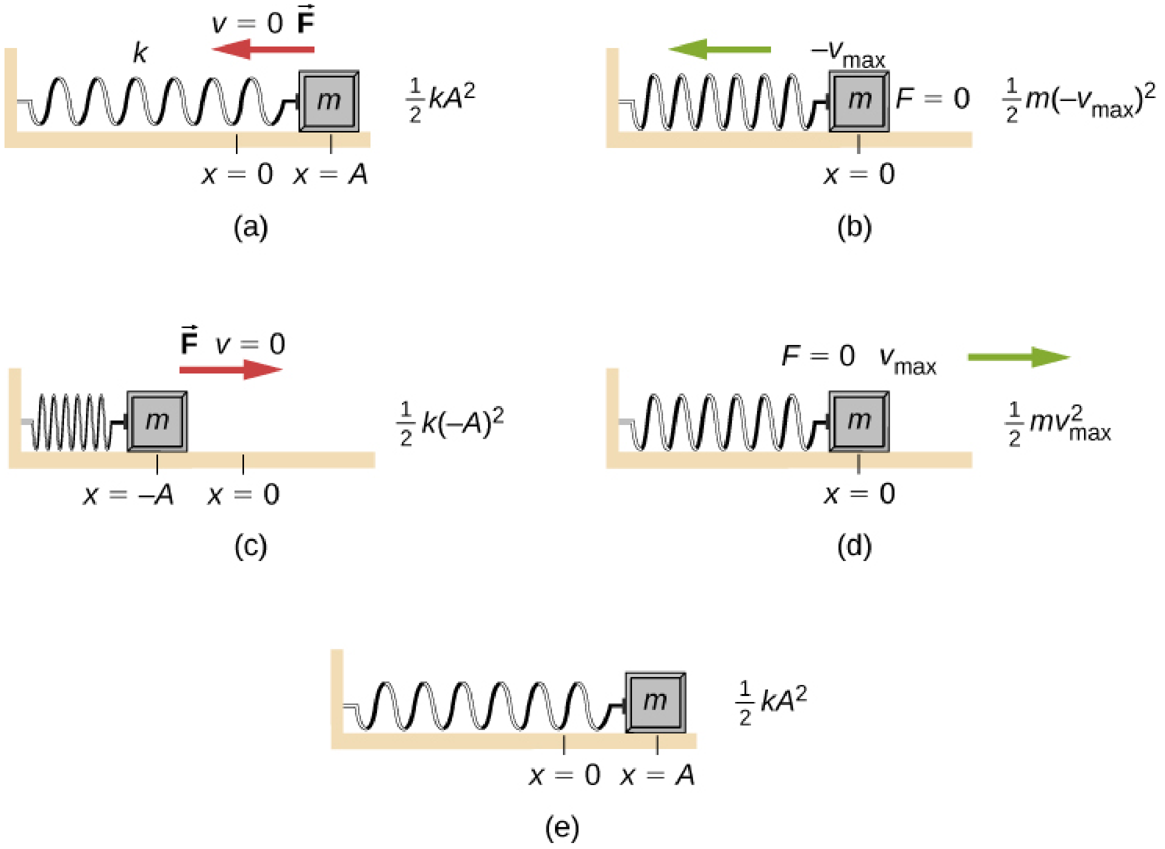
To study the energy of a simple harmonic oscillator, we need to consider all the forms of energy. Consider the example of a block attached to a spring, placed on a frictionless surface, oscillating in SHM. The potential energy stored in the deformation of the spring is
Equation:

$$U=\frac{1}{2}kx^2.$$

In a simple harmonic oscillator, the energy oscillates between kinetic energy of the mass $K=\frac{1}{2}mv^2$ and potential energy $U=\frac{1}{2}kx^2$ stored in the spring. In the SHM of the mass and spring system, there are no dissipative forces, so the total energy is the sum of the potential energy and kinetic energy. In this section, we consider the conservation of energy of the system.

The concepts examined are valid for all simple harmonic oscillators, including those where the gravitational force plays a role.

Consider [link], which shows an oscillating block attached to a spring. In the case of undamped SHM, the energy oscillates back and forth between kinetic and potential, going completely from one form of energy to the other as the system oscillates. So for the simple example of an object on a frictionless surface attached to a spring, the motion starts with all of the energy stored in the spring as **elastic potential energy**. As the object starts to move, the elastic potential energy is converted into kinetic energy, becoming entirely kinetic energy at the equilibrium position. The energy is then converted back into elastic potential energy by the spring as it is stretched or compressed. The velocity becomes zero when the kinetic energy is completely converted, and this cycle then repeats. Understanding the conservation of energy in these cycles will provide extra insight here and in later applications of SHM, such as alternating circuits.



The transformation of energy in SHM for an object attached to a spring on a frictionless surface. (a) When the mass is at the position $x = +A$, all the energy is stored as potential energy in the spring $U = \frac{1}{2}kA^2$. The kinetic energy is equal to zero because the velocity of the mass is zero. (b) As the mass moves toward $x = -A$, the mass crosses the position $x = 0$. At this point, the spring is neither extended nor compressed, so the potential energy stored in the spring is zero. At $x = 0$, the total energy is all kinetic energy where $K = \frac{1}{2}m(-v_{\text{max}})^2$. (c) The mass continues to move until it reaches $x=-A$ where the mass stops and starts moving toward $x=+A$. At the position $x=-A$, the total energy is stored as potential energy in the compressed $U=\frac{1}{2}k(-A)^2$ and the kinetic energy is zero. (d) As the mass passes through the position $x=0$, the kinetic energy is $K=\frac{1}{2}mv_{\text{max}}^2$ and the potential energy stored in the spring is zero. (e) The mass returns to the position $x=+A$, where $K=0$ and $U=\frac{1}{2}kA^2$.

Consider [link], which shows the energy at specific points on the periodic motion. While staying constant, the energy oscillates between the kinetic energy of the block and the potential energy stored in the spring:

Equation:

$$E_{\text{Total}}=U+K=\frac{1}{2}kx^2+\frac{1}{2}mv^2.$$

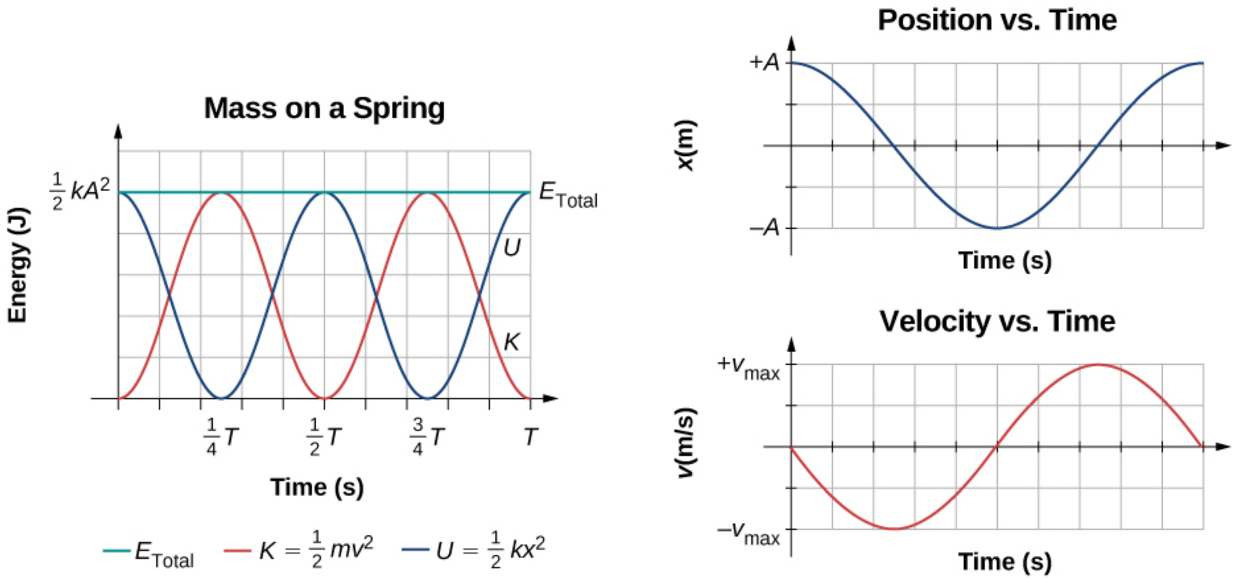
The motion of the block on a spring in SHM is defined by the position $x\left(t\right)=A\text{cos}\left(\omega t+\phi\right)$ with a velocity of $v\left(t\right)=-A\omega\text{sin}\left(\omega t+\phi\right)$. Using these equations, the trigonometric identity $\text{cos}^2\theta+\text{sin}^2\theta=1$ and $\omega=\sqrt{\frac{k}{m}}$, we can find the total energy of the system:

Equation:

$$\begin{aligned}E_{\text{Total}}&=\tfrac{1}{2}kA^2\text{cos}^2\left(\omega t+\phi\right)+\tfrac{1}{2}mA^2\omega^2\text{sin}^2\left(\omega t+\phi\right)\\&=\tfrac{1}{2}kA^2\text{cos}^2\left(\omega t+\phi\right)+\tfrac{1}{2}mA^2\left(\tfrac{k}{m}\right)\text{sin}^2\left(\omega t+\phi\right)\\&=\tfrac{1}{2}kA^2\text{cos}^2\left(\omega t+\phi\right)+\tfrac{1}{2}kA^2\text{sin}^2\left(\omega t+\phi\right)\\&=\tfrac{1}{2}kA^2\left(\text{cos}^2\left(\omega t+\phi\right)+\text{sin}^2\left(\omega t+\phi\right)\right)\\&=\tfrac{1}{2}kA^2.\end{aligned}$$

The total energy of the system of a block and a spring is equal to the sum of the potential energy stored in the spring plus the kinetic energy of the block and is proportional to the square of the amplitude $E_{\text{Total}}=(1/2)kA^2$. The total energy of the system is constant.

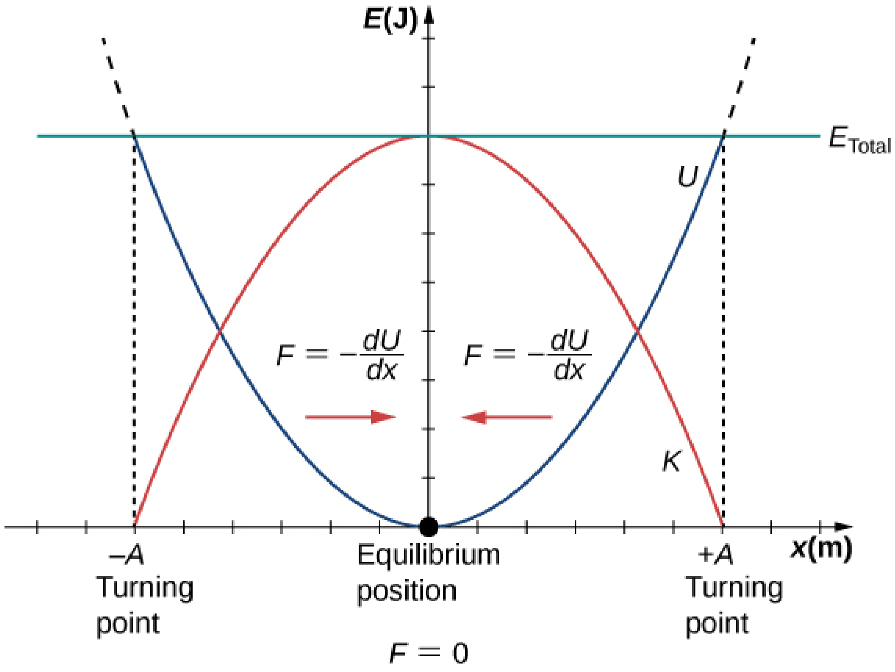
A closer look at the energy of the system shows that the kinetic energy oscillates like a sine-squared function, while the potential energy oscillates like a cosine-squared function. However, the total energy for the system is constant and is proportional to the amplitude squared. [link] shows a plot of the potential, kinetic, and total energies of the block and spring system as a function of time. Also plotted are the position and velocity as a function of time. Before time $t=0.0\,\text{s}$, the block is attached to the spring and placed at the equilibrium position. Work is done on the block by applying an external force, pulling it out to a position of $x=+A$. The system now has potential energy stored in the spring. At time $t=0.00\,\text{s}$, the position of the block is equal to the amplitude, the potential energy stored in the spring is equal to $U=\frac{1}{2}kA^2$, and the force on the block is maximum and points in the negative *x*-direction $\left(F_S=-kA\right)$. The velocity and kinetic energy of the block are zero at time $t=0.00\,\text{s}$. At time $t=0.00\,\text{s}$, the block is released from rest.



Graph of the kinetic energy, potential energy, and total energy of a block oscillating on a spring in SHM. Also shown are the graphs of position versus time and velocity versus time. The total energy remains constant, but the energy oscillates between kinetic energy and potential energy. When the kinetic energy is maximum, the potential energy is zero. This occurs when the velocity is maximum and the mass is at the equilibrium position. The potential energy is maximum when the speed is zero. The total energy is the sum of the kinetic energy plus the potential energy and it is constant.

Oscillations About an Equilibrium Position

We have just considered the energy of SHM as a function of time. Another interesting view of the simple harmonic oscillator is to consider the energy as a function of position. [link] shows a graph of the energy versus position of a system undergoing SHM.



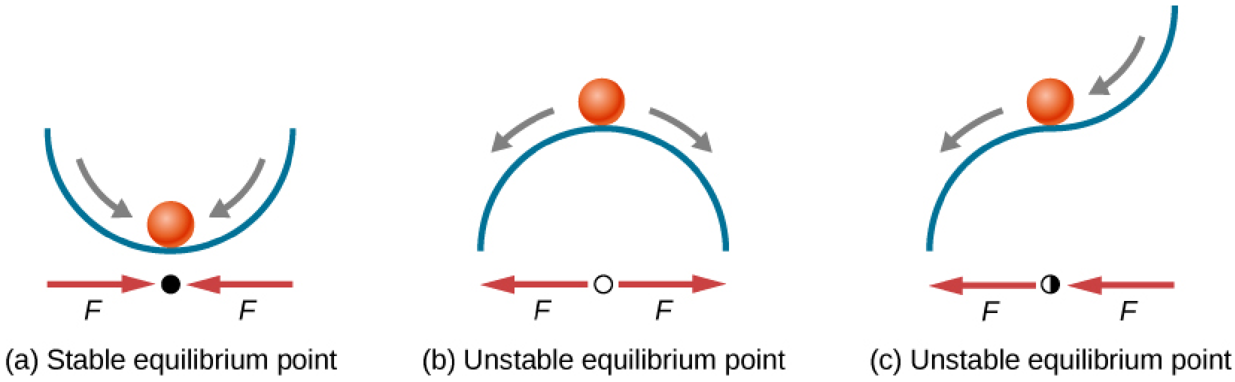
A graph of the kinetic energy (red), potential energy (blue), and total energy (green) of a simple harmonic oscillator. The force is equal to $F = -\frac{dU}{dx}$. The equilibrium position is shown as a black dot and is the point where the force is equal to zero. The force is positive when $x < 0$, negative when $x > 0$, and equal to zero when $x = 0$.

The potential energy curve in [link] resembles a bowl. When a marble is placed in a bowl, it settles to the equilibrium position at the lowest point of the bowl $(x = 0)$. This happens because a **restoring force** points toward the equilibrium point. This equilibrium point is sometimes referred to as a *fixed point*. When the marble is disturbed to a different position $(x = +A)$, the marble oscillates around the equilibrium position. Looking back at the graph of potential energy, the force can be found by looking at the slope of the potential energy graph $\left(F = -\frac{dU}{dx}\right)$. Since the force on either side of the fixed point points back toward the equilibrium point, the equilibrium point is called a **stable equilibrium point**. The points $x = A$ and $x = -A$ are called the turning points. (See Potential Energy and Conservation of Energy.)

Stability is an important concept. If an equilibrium point is stable, a slight disturbance of an object that is initially at the stable equilibrium point will cause the object to oscillate around that point. The stable equilibrium point occurs because the force on either side is directed toward it. For an unstable equilibrium point, if the object is disturbed slightly, it does not return to the equilibrium point.

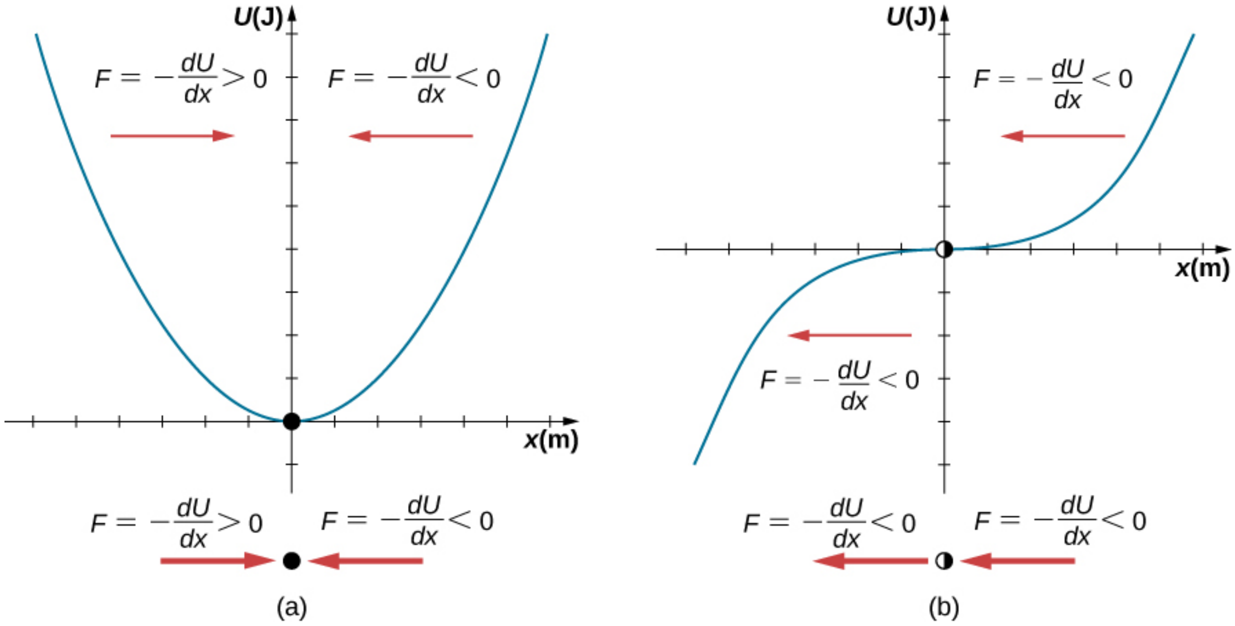
Consider the marble in the bowl example. If the bowl is right-side up, the marble, if disturbed slightly, will oscillate around the stable equilibrium point. If the bowl is turned upside down, the marble can be balanced on the top, at the equilibrium point where the net force is zero. However, if the marble is disturbed slightly, it will not return to the equilibrium point, but will instead roll off the bowl. The reason is that the force on either side of the equilibrium point is directed away from that point. This point is an unstable equilibrium point.

[link] shows three conditions. The first is a stable equilibrium point (a), the second is an unstable equilibrium point (b), and the last is also an unstable equilibrium point (c), because the force on only one side points toward the equilibrium point.


F
F
(a) Stable equilibrium point
F
F
(b) Unstable equilibrium point
F
F
(c) Unstable equilibrium point

Examples of equilibrium points. (a) Stable equilibrium point; (b) unstable equilibrium point; (c) unstable equilibrium point (sometimes referred to as a half-stable equilibrium point).

The process of determining whether an equilibrium point is stable or unstable can be formalized. Consider the potential energy curves shown in [link]. The force can be found by analyzing the slope of the graph. The force is $F=-\frac{dU}{dx}$. In (a), the fixed point is at $x=0.00\,\text{m}$. When $x<0.00\,\text{m}$, the force is positive. When $x>0.00\,\text{m}$, the force is negative. This is a stable point. In (b), the fixed point is at $x=0.00\,\text{m}$. When $x<0.00\,\text{m}$, the force is negative. When $x>0.00\,\text{m}$, the force is also negative. This is an unstable point.



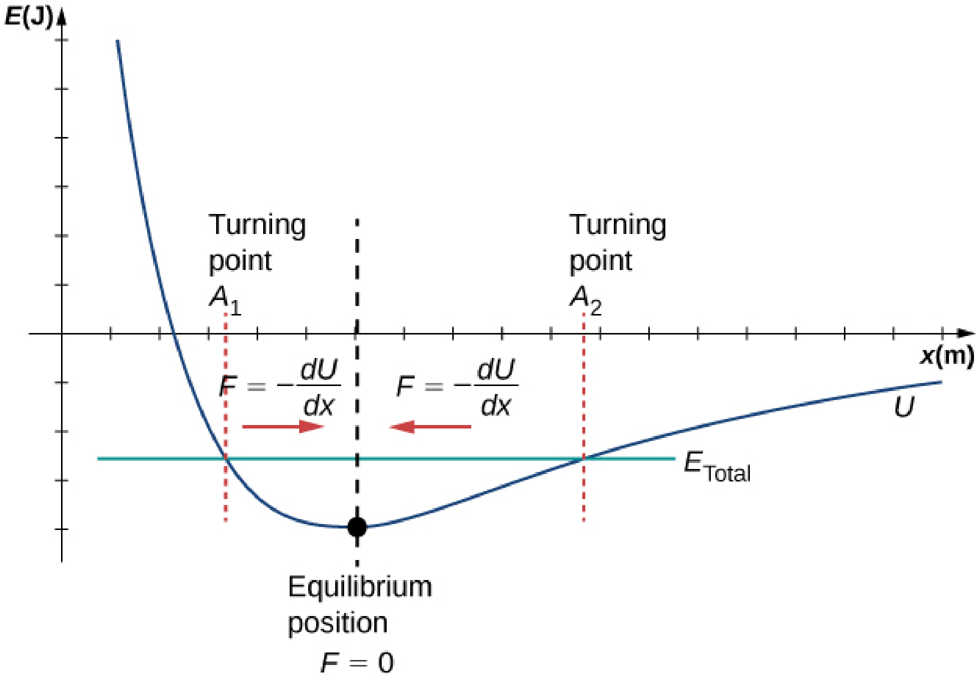
Two examples of a potential energy function. The force at a position is equal to the negative of the slope of the graph at that position. (a) A potential energy function with a stable equilibrium point. (b) A potential energy function with an unstable equilibrium point. This point is sometimes called half-stable because the force on one side points toward the fixed point.

A practical application of the concept of stable equilibrium points is the force between two neutral atoms in a molecule. If two molecules are in close proximity, separated by a few atomic diameters, they can experience an attractive force. If the molecules move close enough so that the electron shells of the other electrons overlap, the force between the molecules becomes repulsive. The attractive force between the two atoms may cause the atoms to form a molecule. The force between the two molecules is not a linear force and cannot be modeled simply as two masses separated by a spring, but the atoms of the molecule can oscillate around an equilibrium point when displaced a small amount from the equilibrium position. The atoms oscillate due the attractive force and repulsive force between the two atoms.

Consider one example of the interaction between two atoms known as the van Der Waals interaction. It is beyond the scope of this chapter to discuss in depth the interactions of the two atoms, but the oscillations of the atoms can be examined by considering one example of a model of the potential energy of the system. One suggestion to model the potential energy of this molecule is with the Lennard-Jones 6-12 potential:
Equation:

$$U\left(x\right)=4\varepsilon\left[{\left(\frac{\sigma}{x}\right)}^{12}-{\left(\frac{\sigma}{x}\right)}^{6}\right].$$

A graph of this function is shown in [link]. The two parameters ε and σ are found experimentally.



The Lennard-Jones potential energy function for a system of two neutral atoms. If the energy is below some maximum energy, the system oscillates near the equilibrium position between the two turning points.

From the graph, you can see that there is a potential energy well, which has some similarities to the potential energy well of the potential energy function of the simple harmonic oscillator discussed in [link]. The Lennard-Jones potential has a stable equilibrium point where the potential energy is minimum and the force on either side of the equilibrium point points toward equilibrium point. Note that unlike the simple harmonic oscillator, the potential well of the Lennard-Jones potential is not symmetric. This is due to the fact that the force between the atoms is not a Hooke's law force and is not linear. The atoms can still oscillate around the equilibrium position x_{min} because when $x < x_{\text{min}}$, the force is positive; when $x > x_{\text{min}}$, the force is negative. Notice that as *x* approaches zero, the slope is quite steep and negative, which means that the force is large and positive. This suggests that it takes a large force to try to push the atoms close together. As *x* becomes increasingly large, the slope becomes less steep and the force is smaller and negative. This suggests that if given a large enough energy, the atoms can be separated.

If you are interested in this interaction, find the force between the molecules by taking the derivative of the potential energy function. You will see immediately that the force does not resemble a Hooke's law force $(F = -kx)$, but if you are familiar with the binomial theorem:

Equation:

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots,$$

the force can be approximated by a Hooke’s law force.

Velocity and Energy Conservation

Getting back to the system of a block and a spring in [link], once the block is released from rest, it begins to move in the negative direction toward the equilibrium position. The potential energy decreases and the magnitude of the velocity and the kinetic energy increase. At time $t = T/4$, the block reaches the equilibrium position $x = 0.00\ \text{m}$, where the force on the block and the potential energy are zero. At the equilibrium position, the block reaches a negative velocity with a magnitude equal to the maximum velocity $v = -A\omega$. The kinetic energy is maximum and equal to $K = \frac{1}{2}mv^2 = \frac{1}{2}mA^2\omega^2 = \frac{1}{2}kA^2$. At this point, the force on the block is zero, but momentum carries the block, and it continues in the negative direction toward $x = -A$. As the block continues to move, the force on it acts in the positive direction and the magnitude of the velocity and kinetic energy decrease. The potential energy increases as the spring compresses. At time $t = T/2$, the block reaches $x = -A$. Here the velocity and kinetic energy are equal to zero. The force on the block is $F = +kA$ and the potential energy stored in the spring is $U = \frac{1}{2}kA^2$. During the oscillations, the total energy is constant and equal to the sum of the potential energy and the kinetic energy of the system,

Note:
Equation:

$$E_{\text{Total}} = \frac{1}{2}kx^2 + \frac{1}{2}mv^2 = \frac{1}{2}kA^2.$$

The equation for the energy associated with SHM can be solved to find the magnitude of the velocity at any position:

Note:
Equation:

$$|v| = \sqrt{\frac{k}{m}(A^2 - x^2)}.$$

The energy in a simple harmonic oscillator is proportional to the square of the amplitude. When considering many forms of oscillations, you will find the energy proportional to the amplitude squared.

Note:
Exercise:

Problem:

Check Your Understanding Why would it hurt more if you snapped your hand with a ruler than with a loose spring, even if the displacement of each system is equal?

Solution:

The ruler is a stiffer system, which carries greater force for the same amount of displacement. The ruler snaps your hand with greater force, which hurts more.

Note:
Exercise:

Problem:

Check Your Understanding Identify one way you could decrease the maximum velocity of a simple harmonic oscillator.

Solution:

You could increase the mass of the object that is oscillating. Other options would be to reduce the amplitude, or use a less stiff spring.

Summary

- The simplest type of oscillations are related to systems that can be described by Hooke's law, $F = -kx$, where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the force constant of the system.
- Elastic potential energy U stored in the deformation of a system that can be described by Hooke's law is given by $U = \frac{1}{2}kx^2$.
- Energy in the simple harmonic oscillator is shared between elastic potential energy and kinetic energy, with the total being constant:
  Equation:

$$E_{\text{Total}} = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 = \frac{1}{2}kA^2 = \text{constant}.$$

- The magnitude of the velocity as a function of position for the simple harmonic oscillator can be found by using
  Equation:

$$|v| = \sqrt{\frac{k}{m}(A^2 - x^2)}.$$

Conceptual Questions

Exercise:

Problem: Describe a system in which elastic potential energy is stored.

Solution:

In a car, elastic potential energy is stored when the shock is extended or compressed. In some running shoes elastic potential energy is stored in the compression of the material of the soles of the running shoes. In pole vaulting, elastic potential energy is stored in the bending of the pole.

Exercise:

Problem:

Explain in terms of energy how dissipative forces such as friction reduce the amplitude of a harmonic oscillator. Also explain how a driving mechanism can compensate. (A pendulum clock is such a system.)

Exercise:

Problem:

The temperature of the atmosphere oscillates from a maximum near noontime and a minimum near sunrise. Would you consider the atmosphere to be in stable or unstable equilibrium?

Solution:

The overall system is stable. There may be times when the stability is interrupted by a storm, but the driving force provided by the sun bring the atmosphere back into a stable pattern.

Problems

Exercise:

Problem:

Fish are hung on a spring scale to determine their mass. (a) What is the force constant of the spring in such a scale if it the spring stretches 8.00 cm for a 10.0 kg load? (b) What is the mass of a fish that stretches the spring 5.50 cm? (c) How far apart are the half-kilogram marks on the scale?

Exercise:

Problem:

It is weigh-in time for the local under-85-kg rugby team. The bathroom scale used to assess eligibility can be described by Hooke's law and is depressed 0.75 cm by its maximum load of 120 kg. (a) What is the spring's effective force constant? (b) A player stands on the scales and depresses it by 0.48 cm. Is he eligible to play on this under-85-kg team?

Solution:

a. $1.57 \times 10^5 \text{ N/m}$; b. 77 kg, yes, he is eligible to play

Exercise:

Problem:

One type of BB gun uses a spring-driven plunger to blow the BB from its barrel. (a) Calculate the force constant of its plunger's spring if you must compress it 0.150 m to drive the 0.0500-kg plunger to a top speed of 20.0 m/s. (b) What force must be exerted to compress the spring?

Exercise:

Problem:

When an 80.0-kg man stands on a pogo stick, the spring is compressed 0.120 m. (a) What is the force constant of the spring? (b) Will the spring be compressed more when he hops down the road?

Solution:

a. $6.53 \times 10^3 \text{ N/m}$; b. yes, when the man is at his lowest point in his hopping the spring will be compressed the most

Exercise:

Problem:

A spring has a length of 0.200 m when a 0.300-kg mass hangs from it, and a length of 0.750 m when a 1.95-kg mass hangs from it. (a) What is the force constant of the spring? (b) What is the unloaded length of the spring?

Exercise:

Problem:

The length of nylon rope from which a mountain climber is suspended has an effective force constant of $1.40 \times 10^4\ \text{N/m}$. (a) What is the frequency at which he bounces, given his mass plus and the mass of his equipment are 90.0 kg? (b) How much would this rope stretch to break the climber’s fall if he free-falls 2.00 m before the rope runs out of slack? (*Hint:* Use conservation of energy.) (c) Repeat both parts of this problem in the situation where twice this length of nylon rope is used.

Solution:

a. 1.99 Hz; b. 50.2 cm; c. 0.710 m

Glossary

elastic potential energy
: potential energy stored as a result of deformation of an elastic object, such as the stretching of a spring

restoring force
: force acting in opposition to the force caused by a deformation

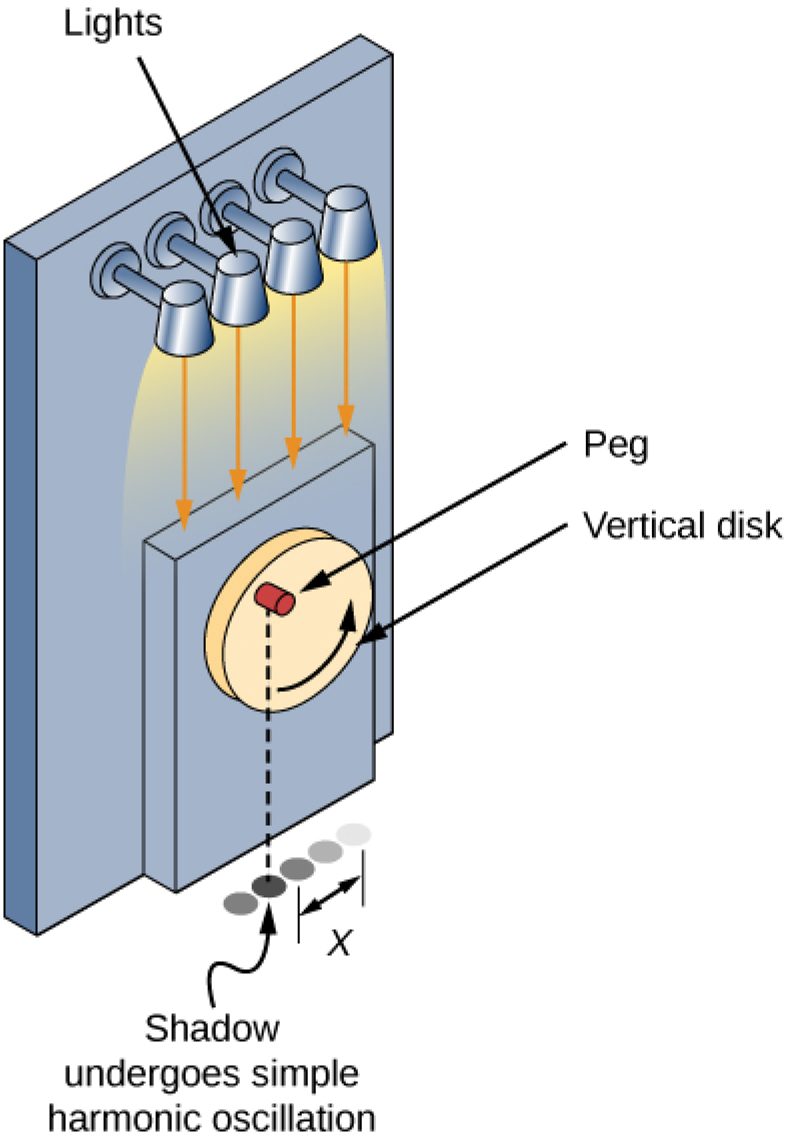
stable equilibrium point
: point where the net force on a system is zero, but a small displacement of the mass will cause a restoring force that points toward the equilibrium point

Comparing Simple Harmonic Motion and Circular Motion

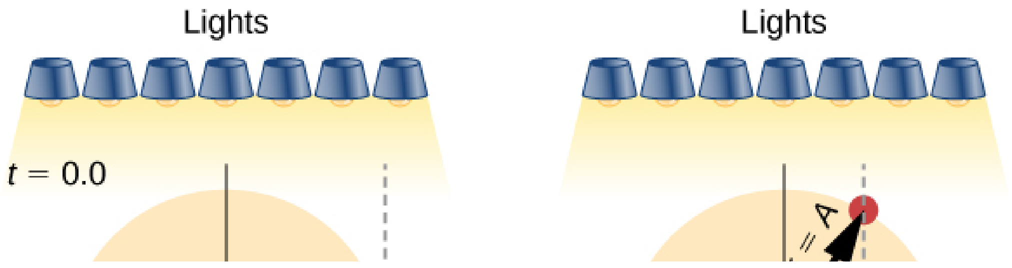
By the end of this section, you will be able to:

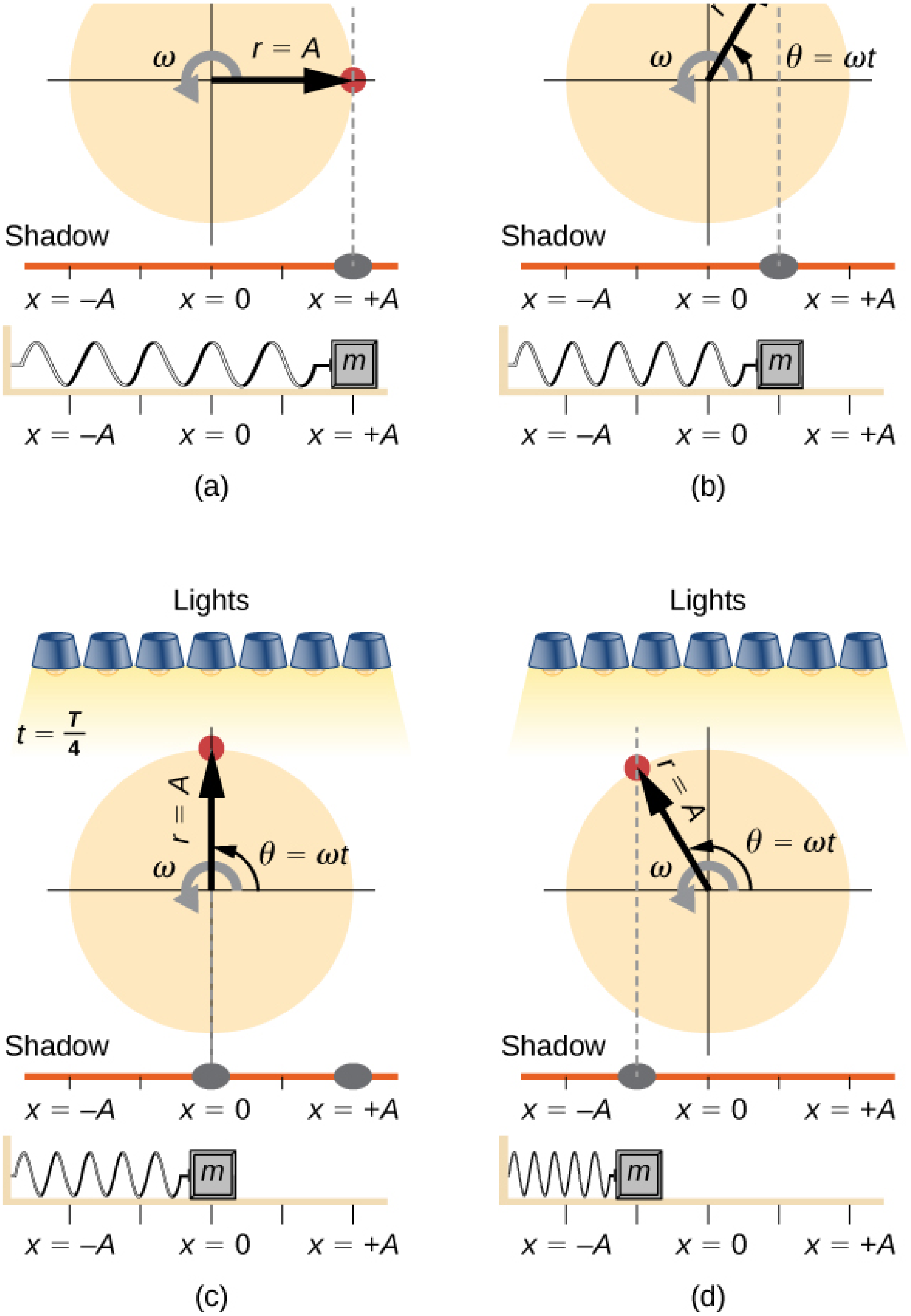
- Describe how the sine and cosine functions relate to the concepts of circular motion
- Describe the connection between simple harmonic motion and circular motion

An easy way to model SHM is by considering uniform circular motion. [link] shows one way of using this method. A peg (a cylinder of wood) is attached to a vertical disk, rotating with a constant angular frequency. [link] shows a side view of the disk and peg. If a lamp is placed above the disk and peg, the peg produces a shadow. Let the disk have a radius of $r = A$ and define the position of the shadow that coincides with the center line of the disk to be $x = 0.00\,m$. As the disk rotates at a constant rate, the shadow oscillates between $x = +A$ and $x = -A$. Now imagine a block on a spring beneath the floor as shown in [link].


Lights
Peg
Vertical disk
X
Shadow undergoes simple harmonic oscillation

SHM can be modeled as rotational motion by looking at the shadow of a peg on a wheel rotating at a constant angular frequency.


Lights
t = 0.0
Lights
A


ω
r = A
θ = ωt
Shadow
x = –A
x = 0
x = +A
m
(a)
(b)
Lights
t = T/4
(c)
(d)

Light shines down on the disk so that the peg makes a shadow. If the disk rotates at just the right angular frequency, the shadow follows the motion of the block on a spring. If there is no energy dissipated due to nonconservative forces, the block and the shadow will oscillate back and forth in unison. In this figure, four snapshots are taken at four different times. (a) The wheel starts at $\theta = 0^o$ and the shadow of the peg is at $x = +A$, representing the mass at position $x = +A$. (b) As the disk rotates through an angle $\theta = \omega t$, the shadow of the peg is between $x = +A$ and $x = 0$. (c) The disk continues to rotate until $\theta = 90^0$, at which the shadow follows the mass to $x = 0$. (d) The disk continues to rotate, the shadow follows the position of the mass.

If the disk turns at the proper angular frequency, the shadow follows along with the block. The position of the shadow can be modeled with the equation

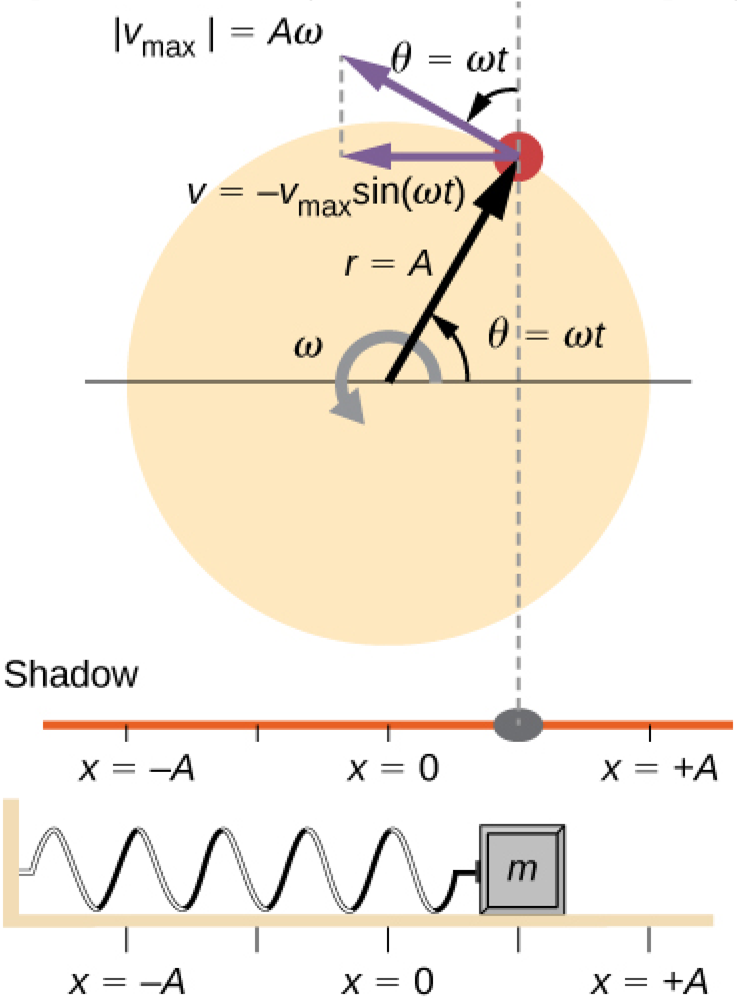
Note:
Equation:

$$x\left(t\right) = A\text{cos}\left(\omega t\right).$$

Recall that the block attached to the spring does not move at a constant velocity. How often does the wheel have to turn to have the peg's shadow always on the block? The disk must turn at a constant angular frequency equal to 2π times the frequency of oscillation $\left(\omega = 2\pi f\right)$.

[link] shows the basic relationship between uniform circular motion and SHM. The peg lies at the tip of the radius, a distance *A* from the center of the disk. The *x*-axis is defined by a line drawn parallel to the ground, cutting the disk in half. The *y*-axis (not shown) is defined by a line perpendicular to the ground, cutting the disk into a left half and a right half. The center of the disk is the point $\left(x = 0, y = 0\right)$. The projection of the

position of the peg onto the fixed x-axis gives the position of the shadow, which undergoes SHM analogous to the system of the block and spring. At the time shown in the figure, the projection has position x and moves to the left with velocity v. The tangential velocity of the peg around the circle equals $\bar{v}_{\text{max}}$ of the block on the spring. The x-component of the velocity is equal to the velocity of the block on the spring.



$|v_{\text{max}}| = A\omega$
$\theta = \omega t$
$v = -v_{\text{max}}\sin(\omega t)$
$r = A$
ω
$\theta = \omega t$
Shadow
$x = -A$
$x = 0$
$x = +A$
m
$x = -A$
$x = 0$
$x = +A$

A peg moving on a circular path with a constant angular velocity ω is undergoing uniform circular motion. Its projection on the x-axis undergoes SHM. Also shown is the velocity of the peg around the circle, v_{max}, and its projection, which is v. Note that these velocities form a similar triangle to the displacement triangle.

We can use [link] to analyze the velocity of the shadow as the disk rotates. The peg moves in a circle with a speed of $v_{\text{max}} = A\omega$. The shadow moves with a velocity equal to the component of the peg's velocity that is parallel to the surface where the shadow is being produced:

Note:
Equation:

$$v = -v_{\text{max}}\sin(\omega t).$$

It follows that the acceleration is

Note:
Equation:

$$a = -a_{\text{max}}\cos(\omega t).$$

Note:
Exercise:

Problem:

Check Your Understanding Identify an object that undergoes uniform circular motion. Describe how you could trace the SHM of this object.

Solution:

A ketchup bottle sits on a lazy Susan in the center of the dinner table. You set it rotating in uniform circular motion. A set of lights shine on the bottle, producing a shadow on the wall.

Summary

- A projection of uniform circular motion undergoes simple harmonic oscillation.
- Consider a circle with a radius *A*, moving at a constant angular speed ω. A point on the edge of the circle moves at a constant tangential speed of $v_{\text{max}} = A\omega$. The projection of the radius onto the *x*-axis is $x\left(t\right) = A\text{cos}\left(\omega t + \phi\right)$, where $\left(\phi\right)$ is the phase shift. The *x*-component of the tangential velocity is $v\left(t\right) = -A\omega\text{sin}\left(\omega t + \phi\right)$.

Conceptual Questions

Exercise:

Problem:

Can this analogy of SHM to circular motion be carried out with an object oscillating on a spring vertically hung from the ceiling? Why or why not? If given the choice, would you prefer to use a sine function or a cosine function to model the motion?

Exercise:

Problem:

If the maximum speed of the mass attached to a spring, oscillating on a frictionless table, was increased, what characteristics of the rotating disk would need to be changed?

Solution:

The maximum speed is equal to $v_{\text{max}} = A\omega$ and the angular frequency is independent of the amplitude, so the amplitude would be affected. The radius of the circle represents the amplitude of the circle, so make the amplitude larger.

Problems

Exercise:

Problem:

The motion of a mass on a spring hung vertically, where the mass oscillates up and down, can also be modeled using the rotating disk. Instead of the lights being placed horizontally along the top and pointing down, place the lights vertically and have the lights shine on the side of the rotating disk. A shadow will be produced on a nearby wall, and will move up and down. Write the equations of motion for the shadow taking the position at $t = 0.0\,\text{s}$ to be $y = 0.0\,\text{m}$ with the mass moving in the positive *y*-direction.

Exercise:

Problem:

(a) A novelty clock has a 0.0100-kg-mass object bouncing on a spring that has a force constant of 1.25 N/m. What is the maximum velocity of the object if the object bounces 3.00 cm above and below its equilibrium position? (b) How many joules of kinetic energy does the object have at its maximum velocity?

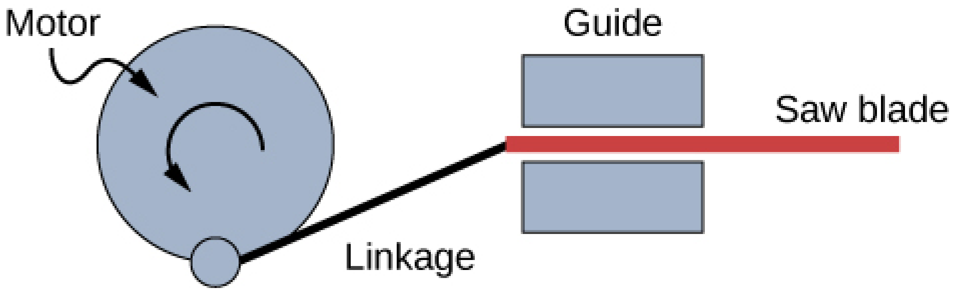
Solution:

a. 0.335 m/s; b. 5.61×10^{-4} J

Exercise:

Problem:

Reciprocating motion uses the rotation of a motor to produce linear motion up and down or back and forth. This is how a reciprocating saw operates, as shown below.



If the motor rotates at 60 Hz and has a radius of 3.0 cm, estimate the maximum speed of the saw blade as it moves up and down. This design is known as a scotch yoke.

Exercise:

Problem:

A student stands on the edge of a merry-go-round which rotates five times a minute and has a radius of two meters one evening as the sun is setting. The student produces a shadow on the nearby building. (a) Write an equation for the position of the shadow. (b) Write an equation for the velocity of the shadow.

Solution:

a. $x\left(t\right)=2\;\text{m}\text{cos}\left(0.52{\text{s}}^{-1}t\right)$; b. $v\left(t\right)=\left(-1.05\;\text{m/s}\right)\text{sin}\left(0.52{\text{s}}^{-1}t\right)$

Pendulums

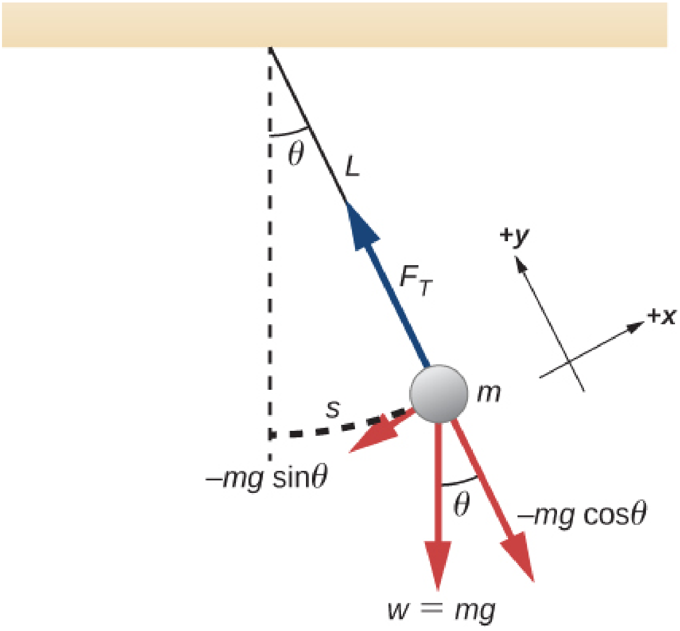
By the end of this section, you will be able to:

- State the forces that act on a simple pendulum
- Determine the angular frequency, frequency, and period of a simple pendulum in terms of the length of the pendulum and the acceleration due to gravity
- Define the period for a physical pendulum
- Define the period for a torsional pendulum

Pendulums are in common usage. Grandfather clocks use a pendulum to keep time and a pendulum can be used to measure the acceleration due to gravity. For small displacements, a pendulum is a simple harmonic oscillator.

The Simple Pendulum

A **simple pendulum** is defined to have a point mass, also known as the pendulum bob, which is suspended from a string of length L with negligible mass ([link]). Here, the only forces acting on the bob are the force of gravity (i.e., the weight of the bob) and tension from the string. The mass of the string is assumed to be negligible as compared to the mass of the bob.



A simple pendulum has a small-diameter bob and a string that has a very small mass but is strong enough not to stretch appreciably. The linear displacement from equilibrium is s, the length of the arc. Also shown are the forces on the bob, which result in a net force of $-mg\sin\theta$ toward the equilibrium position—

that is, a restoring force.

Consider the torque on the pendulum. The force providing the restoring torque is the component of the weight of the pendulum bob that acts along the arc length. The torque is the length of the string L times the component of the net force that is perpendicular to the radius of the arc. The minus sign indicates the torque acts in the opposite direction of the angular displacement:
Equation:

$$\begin{aligned} \tau &= -L\left(mg\sin\theta\right); \\ I\alpha &= -L\left(mg\sin\theta\right); \\ I\frac{d^2\theta}{dt^2} &= -L\left(mg\sin\theta\right); \\ mL^2\frac{d^2\theta}{dt^2} &= -L\left(mg\sin\theta\right); \\ \frac{d^2\theta}{dt^2} &= -\frac{g}{L}\sin\theta. \end{aligned}$$

The solution to this differential equation involves advanced calculus, and is beyond the scope of this text. But note that for small angles (less than 15 degrees), $\sin\theta$ and θ differ by less than 1%, so we can use the small angle approximation $\sin\theta \approx \theta$. The angle θ describes the position of the pendulum. Using the small angle approximation gives an approximate solution for small angles,

Note:
Equation:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\theta.$$

Because this equation has the same form as the equation for SHM, the solution is easy to find. The angular frequency is

Note:
Equation:

$$\omega = \sqrt{\frac{g}{L}}$$

and the period is

Note:
Equation:

$$T = 2\pi\sqrt{\frac{L}{g}}.$$

The period of a simple pendulum depends on its length and the acceleration due to gravity. The period is completely independent of other factors, such as mass and the maximum displacement. As with simple harmonic oscillators, the period T for a pendulum is nearly independent of amplitude, especially if θ is less than about 15°. Even simple pendulum clocks can be finely adjusted and remain accurate.

Note the dependence of T on g. If the length of a pendulum is precisely known, it can actually be used to measure the acceleration due to gravity, as in the following example.

Example:
Measuring Acceleration due to Gravity by the Period of a Pendulum
What is the acceleration due to gravity in a region where a simple pendulum having a length 75.000 cm has a period of 1.7357 s?
Strategy
We are asked to find g given the period T and the length L of a pendulum. We can solve $T = 2\pi\sqrt{\frac{L}{g}}$ for g, assuming only that the angle of deflection is less than 15°.
Solution

1. Square $T = 2\pi\sqrt{\frac{L}{g}}$ and solve for g:
   Equation:

   $$g = 4\pi^2 \frac{L}{T^2}.$$

2. Substitute known values into the new equation:
   Equation:

   $$g = 4\pi^2 \frac{0.75000\ \text{m}}{(1.7357\ \text{s})^2}.$$

3. Calculate to find g:

Equation:

$$g = 9.8281\ \text{m/s}^2.$$

Significance

This method for determining *g* can be very accurate, which is why length and period are given to five digits in this example. For the precision of the approximation $\sin\theta \approx \theta$ to be better than the precision of the pendulum length and period, the maximum displacement angle should be kept below about 0.5°.

Note:

Exercise:

Problem:

Check Your Understanding An engineer builds two simple pendulums. Both are suspended from small wires secured to the ceiling of a room. Each pendulum hovers 2 cm above the floor. Pendulum 1 has a bob with a mass of 10 kg. Pendulum 2 has a bob with a mass of 100 kg. Describe how the motion of the pendulums will differ if the bobs are both displaced by 12°.

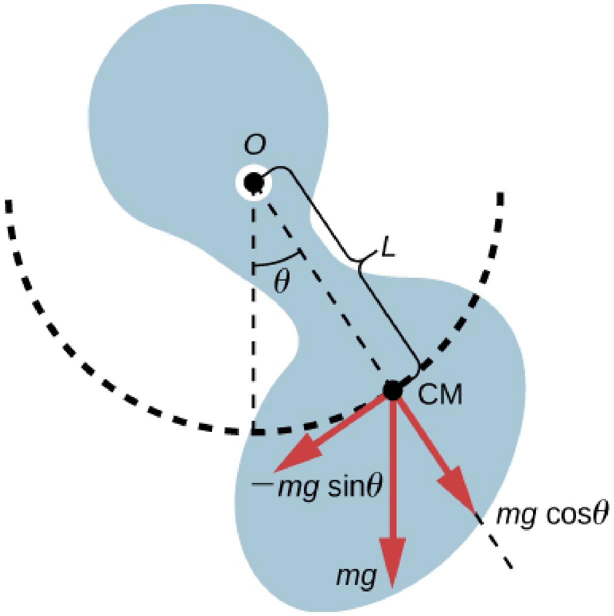
Solution:

The movement of the pendulums will not differ at all because the mass of the bob has no effect on the motion of a simple pendulum. The pendulums are only affected by the period (which is related to the pendulum's length) and by the acceleration due to gravity.

Physical Pendulum

Any object can oscillate like a pendulum. Consider a coffee mug hanging on a hook in the pantry. If the mug gets knocked, it oscillates back and forth like a pendulum until the oscillations die out. We have described a simple pendulum as a point mass and a string. A **physical pendulum** is any object whose oscillations are similar to those of the simple pendulum, but cannot be modeled as a point mass on a string, and the mass distribution must be included into the equation of motion.

As for the simple pendulum, the restoring force of the physical pendulum is the force of gravity. With the simple pendulum, the force of gravity acts on the center of the pendulum bob. In the case of the physical pendulum, the force of gravity acts on the center of mass (CM) of an object. The object oscillates about a point *O*. Consider an object of a generic shape as shown in [link].



A physical pendulum is any object that oscillates as a pendulum, but cannot be modeled as a point mass on a string. The force of gravity acts on the center of mass (CM) and provides the restoring force that causes the object to oscillate. The minus sign on the component of the weight that provides the restoring force is present because the force acts in the opposite direction of the increasing angle θ.

When a physical pendulum is hanging from a point but is free to rotate, it rotates because of the torque applied at the CM, produced by the component of the object's weight that acts tangent to the motion of the CM. Taking the counterclockwise direction to be positive, the component of the gravitational force that acts tangent to the motion is $-mg\sin\theta$. The minus sign is the result of the restoring force acting in the opposite direction of the increasing angle. Recall that the torque is equal to $\vec{\tau} = \vec{\mathbf{r}} \times \vec{\mathbf{F}}$. The magnitude of the torque is equal to the length of the radius arm times the tangential component of the force applied, $|\tau| = rF\sin\theta$. Here, the length L of the radius arm is the distance between the point of rotation and the CM. To analyze the motion, start with the net torque. Like the simple pendulum, consider only small angles so that $\sin\theta \approx \theta$. Recall from Fixed-Axis Rotation on rotation that the net torque is equal to the moment of inertia $I = \int r^2 dm$ times the angular acceleration α, where $\alpha = \frac{d^2\theta}{dt^2}$:

Equation:

$$I\alpha = \tau_{\text{net}} = L\,(-mg)\sin\theta.$$

Using the small angle approximation and rearranging:
Equation:

$$\begin{aligned} I\alpha &= -L\,(mg)\theta; \\ I\frac{d^2\theta}{dt^2} &= -L\,(mg)\theta; \\ \frac{d^2\theta}{dt^2} &= -\left(\frac{mgL}{I}\right)\theta. \end{aligned}$$

Once again, the equation says that the second time derivative of the position (in this case, the angle) equals minus a constant $\left(-\frac{mgL}{I}\right)$ times the position. The solution is
Equation:

$$\theta\,(t) = \Theta\cos\,(\omega t + \phi),$$

where Θ is the maximum angular displacement. The angular frequency is

Note:
Equation:

$$\omega = \sqrt{\frac{mgL}{I}}.$$

The period is therefore

Note:
Equation:

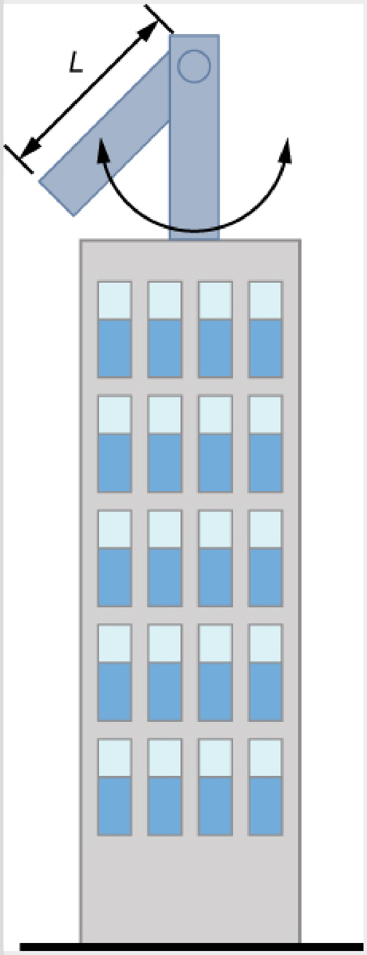
$$T = 2\pi\sqrt{\frac{I}{mgL}}.$$

Note that for a simple pendulum, the moment of inertia is $I = \int r^2 dm = mL^2$ and the period reduces to $T = 2\pi\sqrt{\frac{L}{g}}$.

Example:

Reducing the Swaying of a Skyscraper

In extreme conditions, skyscrapers can sway up to two meters with a frequency of up to 20.00 Hz due to high winds or seismic activity. Several companies have developed physical pendulums that are placed on the top of the skyscrapers. As the skyscraper sways to the right, the pendulum swings to the left, reducing the sway. Assuming the oscillations have a frequency of 0.50 Hz, design a pendulum that consists of a long beam, of constant density, with a mass of 100 metric tons and a pivot point at one end of the beam. What should be the length of the beam?



Strategy

We are asked to find the length of the physical pendulum with a known mass. We first need to find the moment of inertia of the beam. We can then use the equation for the period of a physical pendulum to find the length.

Solution

1. Find the moment of inertia for the CM:
2. Use the parallel axis theorem to find the moment of inertia about the point of rotation:
   Equation:

$$I = I_{\text{CM}} + \frac{L}{4}^2 M = \frac{1}{12}ML^2 + \frac{1}{4}ML^2 = \frac{1}{3}ML^2.$$

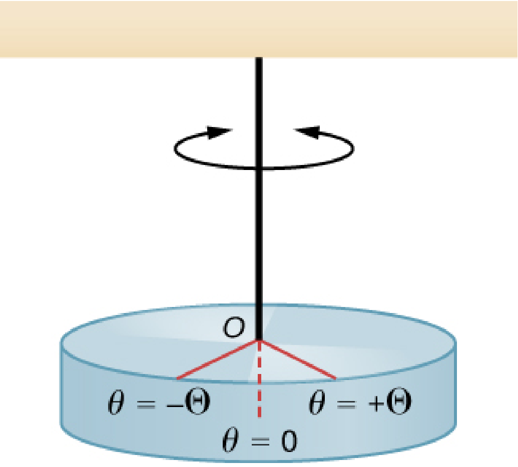
3. The period of a physical pendulum has a period of $T = 2\pi\sqrt{\frac{I}{mgL}}$. Use the moment of inertia to solve for the length L:
   Equation:

$$\begin{aligned} T &= 2\pi\sqrt{\frac{I}{MgL}} = 2\pi\sqrt{\frac{\frac{1}{3}ML^2}{MgL}} = 2\pi\sqrt{\frac{L}{3g}}; \\ L &= 3g\left(\frac{T}{2\pi}\right)^2 = 3\left(9.8\,\frac{\text{m}}{\text{s}^2}\right)\left(\frac{2\,\text{s}}{2\pi}\right)^2 = 2.98\,\text{m}. \end{aligned}$$

Significance
There are many ways to reduce the oscillations, including modifying the shape of the skyscrapers, using multiple physical pendulums, and using tuned-mass dampers.

Torsional Pendulum

A **torsional pendulum** consists of a rigid body suspended by a light wire or spring ([link]). When the body is twisted some small maximum angle (Θ) and released from rest, the body oscillates between $(\theta = +\Theta)$ and $(\theta = -\Theta)$. The restoring torque is supplied by the shearing of the string or wire.



A torsional pendulum consists of a rigid body suspended by a string or wire. The rigid body oscillates between $\theta = +\Theta$ and $\theta = -\Theta$.

The restoring torque can be modeled as being proportional to the angle:

Equation:

$$\tau = -\kappa\theta.$$

The variable kappa (κ) is known as the torsion constant of the wire or string. The minus sign shows that the restoring torque acts in the opposite direction to increasing angular displacement. The net torque is equal to the moment of inertia times the angular acceleration:

Equation:

$$I\frac{d^2\theta}{dt^2} = -\kappa\theta;$$
$$\frac{d^2\theta}{dt^2} = -\frac{\kappa}{I}\theta.$$

This equation says that the second time derivative of the position (in this case, the angle) equals a negative constant times the position. This looks very similar to the equation of motion for the SHM $\frac{d^2x}{dt^2} = -\frac{k}{m}x$, where the period was found to be $T = 2\pi\sqrt{\frac{m}{k}}$. Therefore, the period of the torsional pendulum can be found using

Note:

Equation:

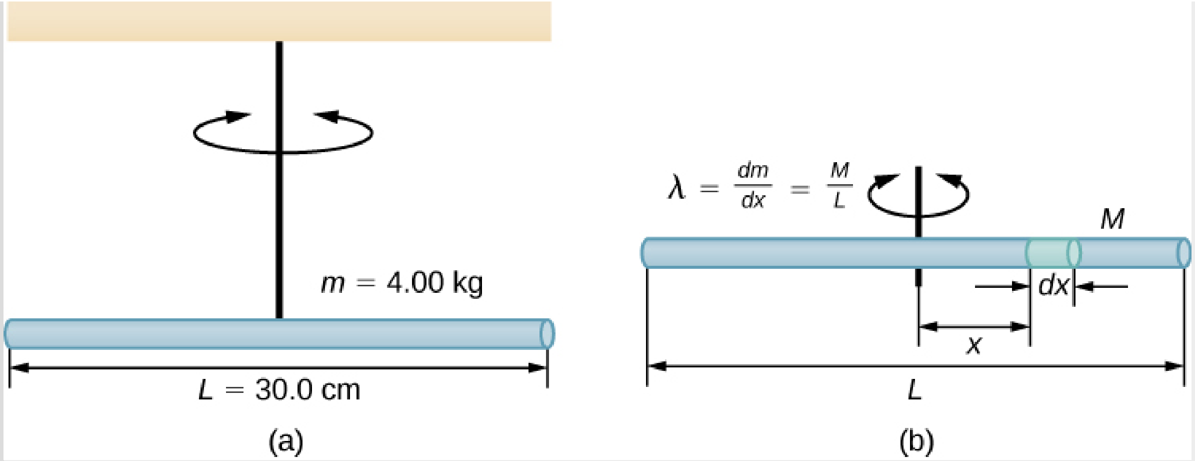
$$T = 2\pi\sqrt{\frac{I}{\kappa}}.$$

The units for the torsion constant are $[\kappa] = \text{N-m} = \left(\text{kg}\frac{\text{m}}{\text{s}^2}\right)\text{m} = \text{kg}\ \frac{\text{m}^2}{\text{s}^2}$ and the units for the moment of inertial are $[I] = \text{kg-m}^2$, which show that the unit for the period is the second.

Example:

Measuring the Torsion Constant of a String

A rod has a length of $l = 0.30\ \text{m}$ and a mass of 4.00 kg. A string is attached to the CM of the rod and the system is hung from the ceiling ([link]). The rod is displaced 10 degrees from the equilibrium position and released from rest. The rod oscillates with a period of 0.5 s. What is the torsion constant κ?



(a) A rod suspended by a string from the ceiling. (b) Finding the rod’s moment of inertia.

Strategy
We are asked to find the torsion constant of the string. We first need to find the moment of inertia.
Solution

1. Find the moment of inertia for the CM:
   Equation:

$$I_{\text{CM}} = \int x^2 dm = \int_{-L/2}^{+L/2} x^2 \lambda dx = \lambda \left[\frac{x^3}{3}\right]_{-L/2}^{+L/2} = \lambda \frac{2L^3}{24} = \left(\frac{M}{L}\right)\frac{2L^3}{24} = \frac{1}{12}ML^2.$$

2. Calculate the torsion constant using the equation for the period:
   Equation:

$$\begin{aligned} T &= 2\pi\sqrt{\tfrac{I}{\kappa}}; \\ \kappa &= I\left(\tfrac{2\pi}{T}\right)^2 = \left(\tfrac{1}{12}ML^2\right)\left(\tfrac{2\pi}{T}\right)^2; \\ &= \left(\tfrac{1}{12}(4.00\ \text{kg})(0.30\ \text{m})^2\right)\left(\tfrac{2\pi}{0.50\ \text{s}}\right)^2 = 4.73\ \text{N} \cdot \text{m}. \end{aligned}$$

Significance
Like the force constant of the system of a block and a spring, the larger the torsion constant, the shorter the period.

Summary

- A mass *m* suspended by a wire of length L and negligible mass is a simple pendulum and undergoes SHM for amplitudes less than about 15°. The period of a simple pendulum is $T = 2\pi\sqrt{\frac{L}{g}}$, where L is the length of the string and *g* is the acceleration due to gravity.
- The period of a physical pendulum $T = 2\pi\sqrt{\frac{I}{mgL}}$ can be found if the moment of inertia is known. The length between the point of rotation and the center of mass is L.
- The period of a torsional pendulum $T = 2\pi\sqrt{\frac{I}{\kappa}}$ can be found if the moment of inertia and torsion constant are known.

Conceptual Questions

Exercise:

Problem:

Pendulum clocks are made to run at the correct rate by adjusting the pendulum’s length. Suppose you move from one city to another where the acceleration due to gravity is slightly greater, taking your pendulum clock with you, will you have to lengthen or shorten the pendulum to keep the correct time, other factors remaining constant? Explain your answer.

Exercise:

Problem:

A pendulum clock works by measuring the period of a pendulum. In the springtime the clock runs with perfect time, but in the summer and winter the length of the pendulum changes. When most materials are heated, they expand. Does the clock run too fast or too slow in the summer? What about the winter?

Solution:

The period of the pendulum is $T = 2\pi\sqrt{L/g}$. In summer, the length increases, and the period increases. If the period should be one second, but period is longer than one second in the summer, it will oscillate fewer than 60 times a minute and clock will run slow. In the winter it will run fast.

Exercise:

Problem:

With the use of a phase shift, the position of an object may be modeled as a cosine or sine function. If given the option, which function would you choose? Assuming that the phase shift is zero, what are the initial conditions of function; that is, the initial position, velocity, and acceleration, when using a sine function? How about when a cosine function is used?

Problems

Exercise:

Problem: What is the length of a pendulum that has a period of 0.500 s?

Exercise:

Problem:

Some people think a pendulum with a period of 1.00 s can be driven with “mental energy” or psycho kinetically, because its period is the same as an average heartbeat. True or not, what is the length of such a pendulum?

Solution:

24.8 cm

Exercise:

Problem: What is the period of a 1.00-m-long pendulum?

Exercise:

Problem:

How long does it take a child on a swing to complete one swing if her center of gravity is 4.00 m below the pivot?

Solution:

4.01 s

Exercise:

Problem: The pendulum on a cuckoo clock is 5.00-cm long. What is its frequency?

Exercise:

Problem:

Two parakeets sit on a swing with their combined CMs 10.0 cm below the pivot. At what frequency do they swing?

Solution:

1.58 s

Exercise:

Problem:

(a) A pendulum that has a period of 3.00000 s and that is located where the acceleration due to gravity is $9.79\ \mathrm{m/s^2}$ is moved to a location where the acceleration due to gravity is $9.82\ \mathrm{m/s^2}$. What is its new period? (b) Explain why so many digits are needed in the value for the period, based on the relation between the period and the acceleration due to gravity.

Exercise:

Problem:

A pendulum with a period of 2.00000 s in one location ($g = 9.80\mathrm{m/s^2}$) is moved to a new location where the period is now 1.99796 s. What is the acceleration due to gravity at its new location?

Solution:

$9.82002\ \mathrm{m/s^2}$

Exercise:

Problem:

(a) What is the effect on the period of a pendulum if you double its length? (b) What is the effect on the period of a pendulum if you decrease its length by 5.00%?

Glossary

physical pendulum	any extended object that swings like a pendulum
simple pendulum	point mass, called a pendulum bob, attached to a near massless string
torsional pendulum	any suspended object that oscillates by twisting its suspension

Damped Oscillations

By the end of this section, you will be able to:

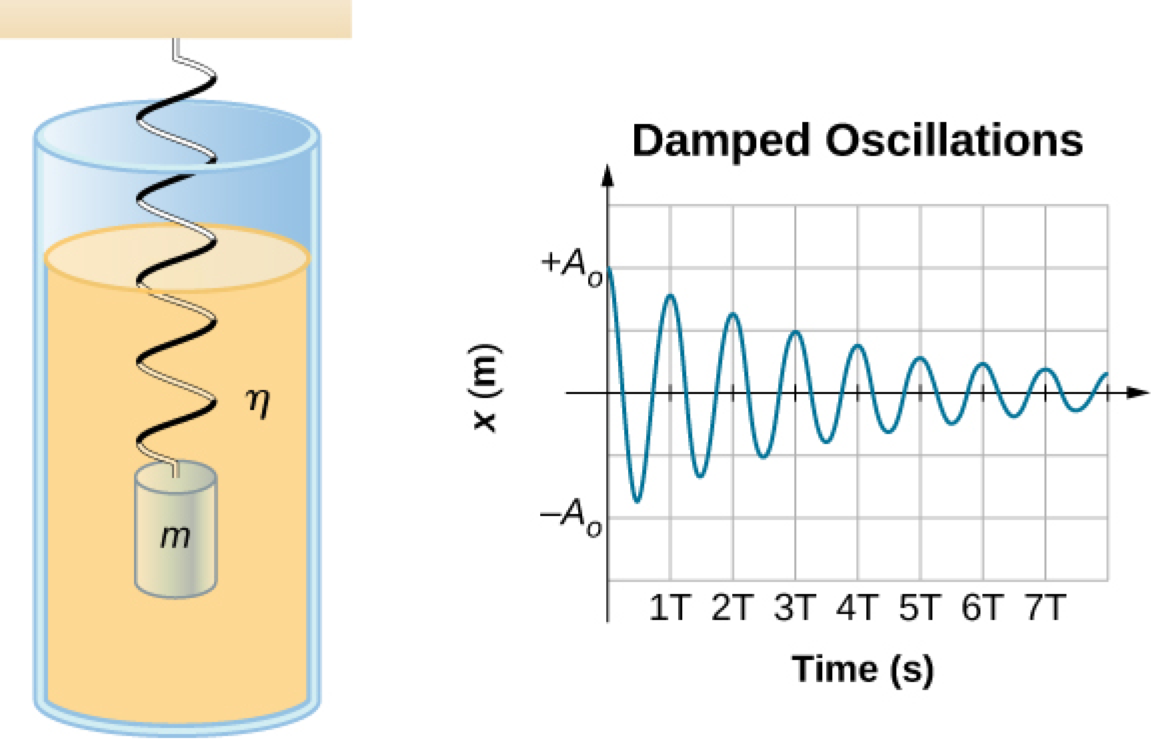
- Describe the motion of damped harmonic motion
- Write the equations of motion for damped harmonic oscillations
- Describe the motion of driven, or forced, damped harmonic motion
- Write the equations of motion for forced, damped harmonic motion

In the real world, oscillations seldom follow true SHM. Friction of some sort usually acts to dampen the motion so it dies away, or needs more force to continue. In this section, we examine some examples of damped harmonic motion and see how to modify the equations of motion to describe this more general case.

A guitar string stops oscillating a few seconds after being plucked. To keep swinging on a playground swing, you must keep pushing ([link]). Although we can often make friction and other nonconservative forces small or negligible, completely undamped motion is rare. In fact, we may even want to damp oscillations, such as with car shock absorbers.

To counteract dampening forces, you need to keep pumping a swing. (credit: Bob Mical)

[link] shows a mass m attached to a spring with a force constant k. The mass is raised to a position A_0, the initial amplitude, and then released. The mass oscillates around the equilibrium position in a fluid with viscosity but the amplitude decreases for each oscillation. For a system that has a small amount of damping, the period and frequency are constant and are nearly the same as for SHM, but the amplitude gradually decreases as shown. This occurs because the non-conservative damping force removes energy from the system, usually in the form of thermal energy.



For a mass on a spring oscillating in a viscous fluid, the period remains constant, but the amplitudes of the oscillations decrease due to the damping caused by the fluid.

Consider the forces acting on the mass. Note that the only contribution of the weight is to change the equilibrium position, as discussed earlier in the chapter. Therefore, the net force is equal to the force of the spring and the damping force (F_D). If the magnitude of the velocity is small, meaning the mass oscillates slowly, the damping force is proportional to the velocity and acts against the direction of motion $(F_D = -bv)$. The net force on the mass is therefore

Equation:

$$ma = -bv - kx.$$

Writing this as a differential equation in *x*, we obtain

Note:

Equation:

$$m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = 0.$$

To determine the solution to this equation, consider the plot of position versus time shown in [link]. The curve resembles a cosine curve oscillating in the envelope of an exponential function $A_0e^{-\alpha t}$ where $\alpha = \frac{b}{2m}$. The solution is

Note:

Equation:

$$x\left(t\right) = A_0e^{-\frac{b}{2m}t}\cos\left(\omega t + \phi\right).$$

It is left as an exercise to prove that this is, in fact, the solution. To prove that it is the right solution, take the first and second derivatives with respect to time and substitute them into [link]. It is found that [link] is the solution if

Equation:

$$\omega = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2}.$$

Recall that the angular frequency of a mass undergoing SHM is equal to the square root of the force constant divided by the mass. This is often referred to as the **natural angular frequency**, which is represented as

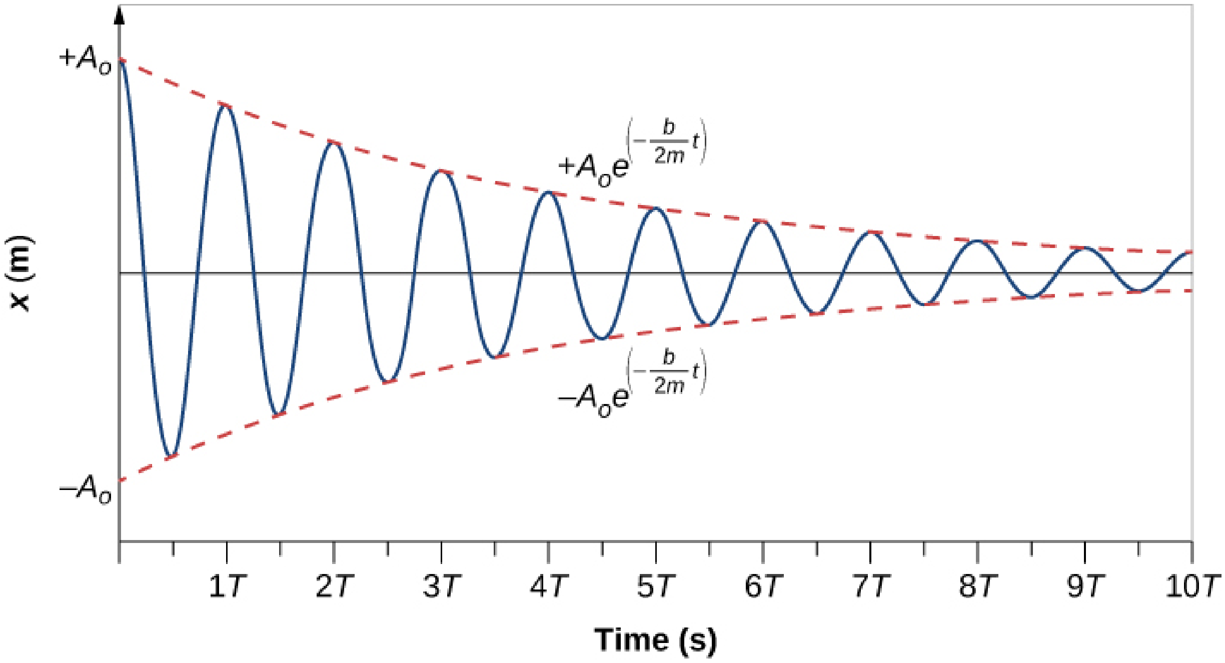
Note:
Equation:

$$\omega_0 = \sqrt{\frac{k}{m}}.$$

The angular frequency for damped harmonic motion becomes

Note:
Equation:

$$\omega = \sqrt{\omega_0^2 - \left(\frac{b}{2m}\right)^2}.$$



Position versus time for the mass oscillating on a spring in a viscous fluid. Notice that the curve appears to be a cosine function inside an exponential envelope.

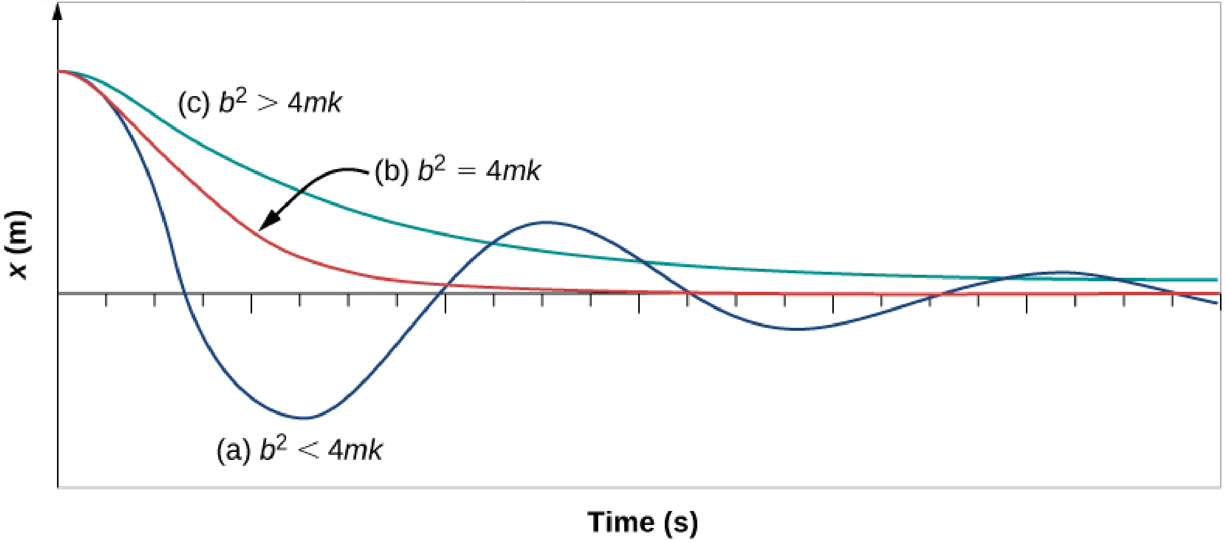
Recall that when we began this description of damped harmonic motion, we stated that the damping must be small. Two questions come to mind. Why must the damping be small? And how small is small? If you gradually *increase* the amount of damping in a system, the period and frequency begin to be affected, because damping opposes and hence slows the back and forth motion. (The net force is smaller in both directions.) If there is very large damping, the system does not even oscillate—it slowly moves toward equilibrium. The angular frequency is equal to

Equation:

$$\omega = \sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2}.$$

As b increases, $\frac{k}{m} - \left(\frac{b}{2m}\right)^2$ becomes smaller and eventually reaches zero when $b = \sqrt{4mk}$. If b becomes any larger, $\frac{k}{m} - \left(\frac{b}{2m}\right)^2$ becomes a negative number and $\sqrt{\frac{k}{m} - \left(\frac{b}{2m}\right)^2}$ is a complex number.

[link] shows the displacement of a harmonic oscillator for different amounts of damping. When the damping constant is small, $b < \sqrt{4mk}$, the system oscillates while the amplitude of the motion decays exponentially. This system is said to be **underdamped**, as in curve (a). Many systems are underdamped, and oscillate while the amplitude decreases exponentially, such as the mass oscillating on a spring. The damping may be quite small, but eventually the mass comes to rest. If the damping constant is $b = \sqrt{4mk}$, the system is said to be **critically damped**, as in curve (b). An example of a critically damped system is the shock absorbers in a car. It is advantageous to have the oscillations decay as fast as possible. Here, the system does not oscillate, but asymptotically approaches the equilibrium condition as quickly as possible. Curve (c) in [link] represents an **overdamped** system where $b > \sqrt{4mk}$. An overdamped system will approach equilibrium over a longer period of time.



The position versus time for three systems consisting of a mass and a spring in a viscous fluid. (a) If the damping is small $\left(b < \sqrt{4mk}\right)$, the mass oscillates, slowly losing amplitude as the energy is dissipated by the non-conservative force(s). The limiting case is (b) where the damping is $\left(b=\sqrt{4mk}\right)$. (c) If the damping is very large $\left(b>\sqrt{4mk}\right)$, the mass does not oscillate when displaced, but attempts to return to the equilibrium position.

Critical damping is often desired, because such a system returns to equilibrium rapidly and remains at equilibrium as well. In addition, a constant force applied to a critically damped system moves the system to a new equilibrium position in the shortest time possible without overshooting or oscillating about the new position.

Note:
Exercise:

Problem:

Check Your Understanding Why are completely undamped harmonic oscillators so rare?

Solution:

Friction often comes into play whenever an object is moving. Friction causes damping in a harmonic oscillator.

Summary

- Damped harmonic oscillators have non-conservative forces that dissipate their energy.
- Critical damping returns the system to equilibrium as fast as possible without overshooting.

- An underdamped system will oscillate through the equilibrium position.
- An overdamped system moves more slowly toward equilibrium than one that is critically damped.

Conceptual Questions

Exercise:

Problem:

Give an example of a damped harmonic oscillator. (They are more common than undamped or simple harmonic oscillators.)

Solution:

A car shock absorber.

Exercise:

Problem:

How would a car bounce after a bump under each of these conditions?

(a) overdamping

(b) underdamping

(c) critical damping

Exercise:

Problem:

Most harmonic oscillators are damped and, if undriven, eventually come to a stop. Why?

Solution:

The second law of thermodynamics states that perpetual motion machines are impossible. Eventually the ordered motion of the system decreases and returns to equilibrium.

Problems

Exercise:

Problem:

The amplitude of a lightly damped oscillator decreases by 3.0 during each cycle. What percentage of the mechanical energy of the oscillator is lost in each cycle?

Solution:

9%

Glossary

critically damped

condition in which the damping of an oscillator causes it to return as quickly as possible to its equilibrium position without oscillating back and forth about this position

natural angular frequency

angular frequency of a system oscillating in SHM

overdamped

condition in which damping of an oscillator causes it to return to equilibrium without oscillating; oscillator moves more slowly toward equilibrium than in the critically damped system

underdamped

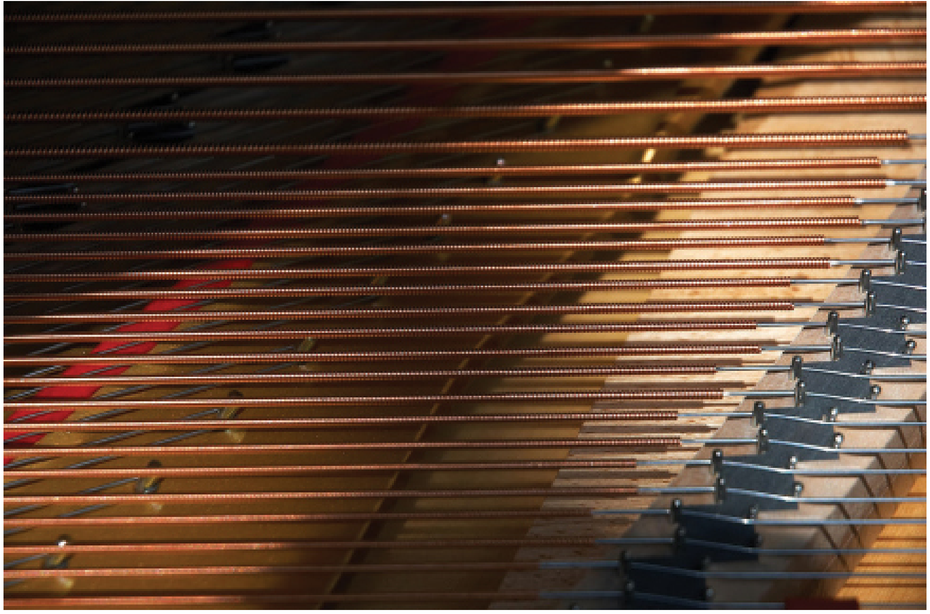
condition in which damping of an oscillator causes the amplitude of oscillations of a damped harmonic oscillator to decrease over time, eventually approaching zero

Forced Oscillations

By the end of this section, you will be able to:

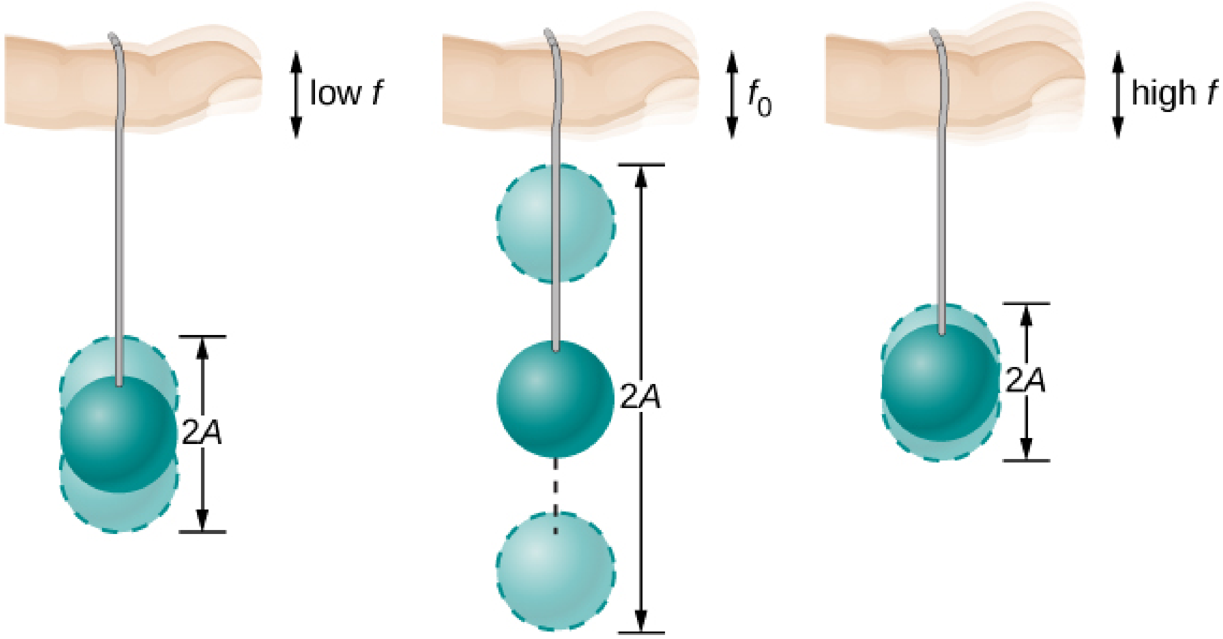
- Define forced oscillations
- List the equations of motion associated with forced oscillations
- Explain the concept of resonance and its impact on the amplitude of an oscillator
- List the characteristics of a system oscillating in resonance

Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings ([link]). It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. This is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate, and oscillate most easily at their natural frequency. In this section, we briefly explore applying a periodic driving force acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. Recall that the natural frequency is the frequency at which a system would oscillate if there were no driving and no damping force.



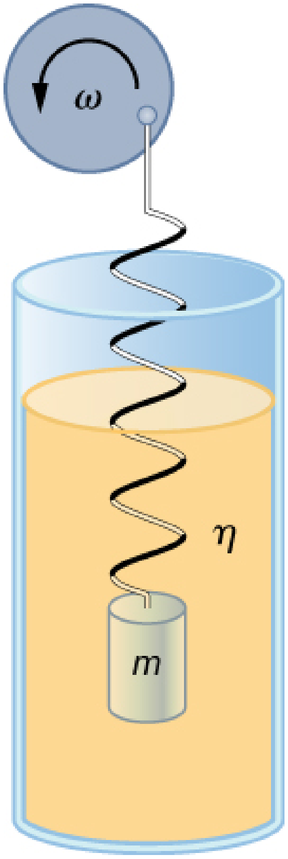
You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings)

Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in [link]. Imagine the finger in the figure is your finger. At first, you hold your finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball follows along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball responds by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called **resonance**. A system being driven at its natural frequency is said to *resonate*. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller until the oscillations nearly disappear, and your finger simply moves up and down with little effect on the ball.



The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f_0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

Consider a simple experiment. Attach a mass *m* to a spring in a viscous fluid, similar to the apparatus discussed in the damped harmonic oscillator. This time, instead of fixing the free end of the spring, attach the free end to a disk that is driven by a variable-speed motor. The motor turns with an angular driving frequency of ω. The rotating disk provides energy to the system by the work done by the driving force $\left(F_{\text{d}} = F_0 \sin\left(\omega t\right)\right)$. The experimental apparatus is shown in [link].



Forced, damped harmonic motion produced by driving a spring and mass with a disk driven by a

variable-
speed motor.

Using Newton's second law ($\vec{\mathbf{F}}_{\text{net}} = m\vec{\mathbf{a}}$), we can analyze the motion of the mass. The resulting equation is similar to the force equation for the damped harmonic oscillator, with the addition of the driving force:

Note:
Equation:

$$-kx - b\frac{dx}{dt} + F_0 \sin(\omega t) = m\frac{d^2x}{dt^2}.$$

When an oscillator is forced with a periodic driving force, the motion may seem chaotic. The motions of the oscillator is known as transients. After the transients die out, the oscillator reaches a steady state, where the motion is periodic. After some time, the steady state solution to this differential equation is

Note:
Equation:

$$x(t) = A\cos(\omega t + \phi).$$

Once again, it is left as an exercise to prove that this equation is a solution. Taking the first and second time derivative of $x(t)$ and substituting them into the force equation shows that $x(t) = A\sin(\omega t + \phi)$ is a solution as long as the amplitude is equal to

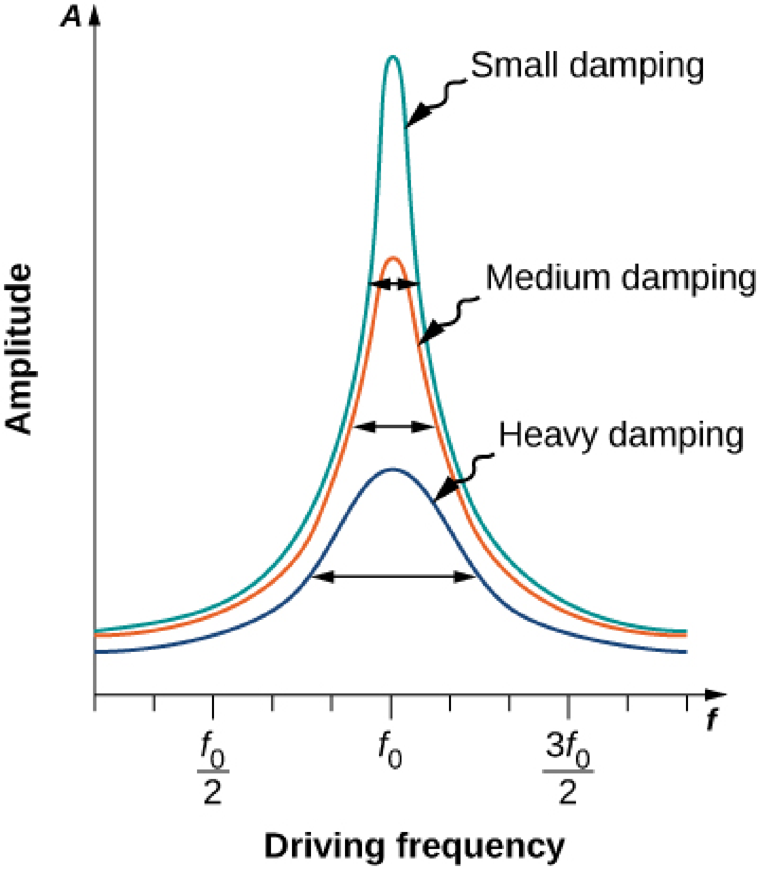
Note:
Equation:

$$A=\frac{F_0}{\sqrt{m^2\left(\omega^2-\omega_0^2\right)^2+b^2\omega^2}}$$

where $\omega_0=\sqrt{\frac{k}{m}}$ is the natural angular frequency of the system of the mass and spring. Recall that the angular frequency, and therefore the frequency, of the motor can be adjusted. Looking at the denominator of the equation for the amplitude, when the driving frequency is much smaller, or much larger, than the natural frequency, the square of the difference of the two angular frequencies $\left(\omega^2-\omega_0^2\right)^2$ is positive and large, making the denominator large, and the result is a small amplitude for the oscillations of the mass. As the frequency of the driving force approaches the natural frequency of the system, the denominator becomes small and the amplitude of the oscillations becomes large. The maximum amplitude results when the frequency of the driving force equals the natural frequency of the system $\left(A_{\text{max}}=\frac{F_0}{b\omega}\right)$.

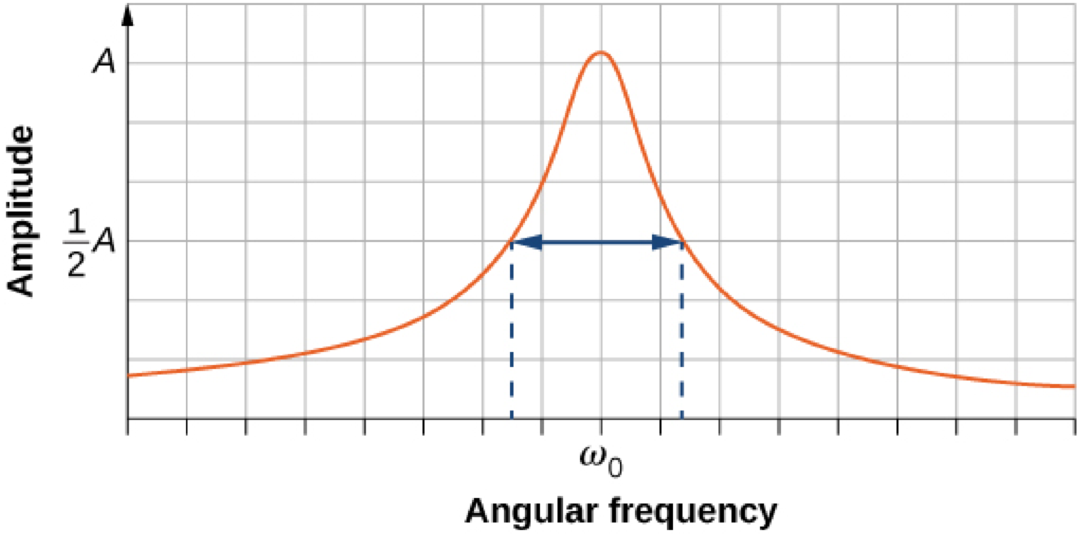
[link] shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. Each of the three curves on the graph represents a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force. Note that since the amplitude grows as the damping decreases, taking this to the limit where there is no damping $(b=0)$, the amplitude becomes infinite.

Note that a small-amplitude driving force can produce a large-amplitude response. This phenomenon is known as resonance. A common example of resonance is a parent pushing a small child on a swing. When the child wants to go higher, the parent does not move back and then, getting a running start, slam into the child, applying a great force in a short interval. Instead, the parent applies small pushes to the child at just the right frequency, and the amplitude of the child’s swings increases.



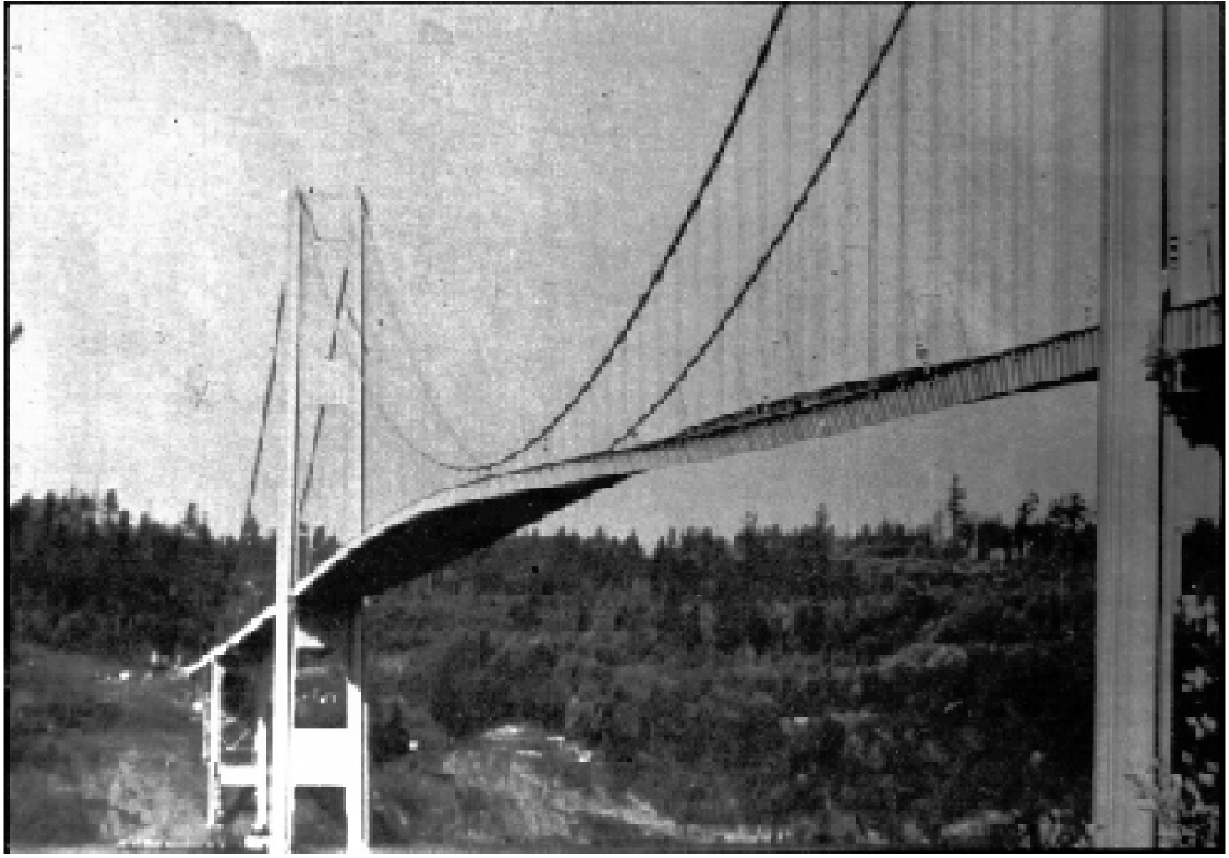
Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting to note that the widths of the resonance curves shown in [link] depend on damping: the less the damping, the narrower the resonance. The consequence is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. For instance, a radio has a circuit that is used to choose a particular radio station. In this case, the forced damped oscillator consists of a resistor, capacitor, and inductor, which will be discussed later in this course. The circuit is “tuned” to pick a particular radio station. Here it is desirable to have the resonance curve be very narrow, to pick out the exact frequency of the radio station chosen. The narrowness of the graph, and the ability to pick out a certain frequency, is known as the quality of the system. The quality is defined as the spread of the angular frequency, or equivalently, the spread in the frequency, at half the maximum amplitude, divided by the natural frequency $\left(Q = \frac{\Delta\omega}{\omega_0}\right)$ as shown in [link]. For a small damping, the quality is approximately equal to $Q \approx \frac{2b}{m}$.



The quality of a system is defined as the spread in the frequencies at half the amplitude divided by the natural frequency.

These features of driven harmonic oscillators apply to a huge variety of systems. For instance, magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei or protons) are made to resonate by incoming radio waves (on the order of 100 MHz). In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance. [link] shows a photograph of a famous example (the Tacoma Narrows bridge) of the destructive effects of a driven harmonic oscillation. The Millennium bridge in London was closed for a short period of time for the same reason while inspections were carried out. Observations lead to modifications being made to the bridge prior to the reopening.



In 1940, the Tacoma Narrows bridge in the state of Washington collapsed. Moderately high, variable cross-winds (much slower than hurricane force winds) drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed. (credit: "PRI's Studio 360"/Flickr)

Note:
Exercise:

Problem:

Check Your Understanding A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Solution:

The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave. With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

Summary

- A system’s natural frequency is the frequency at which the system oscillates if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Key Equations

Relationship between frequency and period	$f = \frac{1}{T}$
Position in SHM with $\phi = 0.00$	$x(t) = A\cos(\omega t)$

General position in SHM	$x\left(t\right)=A\cos\left(\omega t+\phi\right)$
General velocity in SHM	$v\left(t\right)=-A\omega\sin\left(\omega t+\phi\right)$
General acceleration in SHM	$a\left(t\right)=-A\omega^{2}\cos\left(\omega t+\phi\right)$
Maximum displacement (amplitude) of SHM	$x_{\text{max}}=A$
Maximum velocity of SHM	$\lvert v_{\text{max}}\rvert=A\omega$
Maximum acceleration of SHM	$\lvert a_{\text{max}}\rvert=A\omega^{2}$
Angular frequency of a mass-spring system in SHM	$\omega=\sqrt{\frac{k}{m}}$
Period of a mass-spring system in SHM	$T=2\pi\sqrt{\frac{m}{k}}$
Frequency of a mass-spring system in SHM	$f=\frac{1}{2\pi}\sqrt{\frac{k}{m}}$
Energy in a mass-spring system in SHM	$E_{\text{Total}}=\frac{1}{2}kx^{2}+\frac{1}{2}mv^{2}=\frac{1}{2}kA^{2}$
The velocity of the mass in a spring-mass system in SHM	$v=\pm\sqrt{\frac{k}{m}\left(A^{2}-x^{2}\right)}$
The *x*-component of the radius of a rotating disk	$x\left(t\right)=A\cos\left(\omega t+\phi\right)$
The *x*-component of the velocity of the edge of a rotating disk	$v\left(t\right)=-v_{\text{max}}\sin\left(\omega t+\phi\right)$
The *x*-component of the acceleration of the edge of a rotating disk	$a\left(t\right)=-a_{\text{max}}\cos\left(\omega t+\phi\right)$

Force equation for a simple pendulum	$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\theta$
Angular frequency for a simple pendulum	$\omega = \sqrt{\frac{g}{L}}$
Period of a simple pendulum	$T = 2\pi\sqrt{\frac{L}{g}}$
Angular frequency of a physical pendulum	$\omega = \sqrt{\frac{mgL}{I}}$
Period of a physical pendulum	$T = 2\pi\sqrt{\frac{I}{mgL}}$
Period of a torsional pendulum	$T = 2\pi\sqrt{\frac{I}{\kappa}}$
Newton’s second law for harmonic motion	$m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = 0$
Solution for underdamped harmonic motion	$x\left(t\right) = A_0 e^{-\frac{b}{2m}t}\cos\left(\omega t + \phi\right)$
Natural angular frequency of a mass-spring system	$\omega_0 = \sqrt{\frac{k}{m}}$
Angular frequency of underdamped harmonic motion	$\omega = \sqrt{\omega_0^2 - \left(\frac{b}{2m}\right)^2}$
Newton’s second law for forced, damped oscillation	$-kx - b\frac{dx}{dt} + F_o\sin\left(\omega t\right) = m\frac{d^2x}{dt^2}$
Solution to Newton’s second law for forced, damped oscillations	$x\left(t\right) = A\cos\left(\omega t + \phi\right)$
Amplitude of system undergoing forced,	$A = \frac{F_o}{\sqrt{m^2\left(\omega^2 - \omega_o^2\right)^2 + b^2\omega^2}}$

damped oscillations

Conceptual Questions

Exercise:

Problem:

Why are soldiers in general ordered to “route step” (walk out of step) across a bridge?

Exercise:

Problem:

Do you think there is any harmonic motion in the physical world that is not damped harmonic motion? Try to make a list of five examples of undamped harmonic motion and damped harmonic motion. Which list was easier to make?

Solution:

All harmonic motion is damped harmonic motion, but the damping may be negligible. This is due to friction and drag forces. It is easy to come up with five examples of damped motion: (1) A mass oscillating on a hanging on a spring (it eventually comes to rest). (2) Shock absorbers in a car (thankfully they also come to rest). (3) A pendulum is a grandfather clock (weights are added to add energy to the oscillations). (4) A child on a swing (eventually comes to rest unless energy is added by pushing the child). (5) A marble rolling in a bowl (eventually comes to rest). As for the undamped motion, even a mass on a spring in a vacuum will eventually come to rest due to internal forces in the spring. Damping may be negligible, but cannot be eliminated.

Exercise:

Problem:

Some engineers use sound to diagnose performance problems with car engines. Occasionally, a part of the engine is designed that resonates at the frequency of the engine. The unwanted oscillations can cause noise that irritates the driver or could lead to the part failing prematurely. In one case, a part was located that had a length *L* made of a material with a mass *M*. What can be done to correct this problem?

Problems

Exercise:

Problem:

How much energy must the shock absorbers of a 1200-kg car dissipate in order to damp a bounce that initially has a velocity of 0.800 m/s at the equilibrium position? Assume the car returns to its original vertical position.

Exercise:

Problem:

If a car has a suspension system with a force constant of $5.00 \times 10^4\ \text{N/m}$, how much energy must the car’s shocks remove to dampen an oscillation starting with a maximum displacement of 0.0750 m?

Solution:

141 J

Exercise:

Problem:

(a) How much will a spring that has a force constant of 40.0 N/m be stretched by an object with a mass of 0.500 kg when hung motionless from the spring? (b) Calculate the decrease in gravitational potential energy of the 0.500-kg object when it descends this distance. (c) Part of this gravitational energy goes into the spring. Calculate the energy stored in the spring by this stretch, and compare it with the gravitational potential energy. Explain where the rest of the energy might go.

Exercise:

Problem:

Suppose you have a 0.750-kg object on a horizontal surface connected to a spring that has a force constant of 150 N/m. There is simple friction between the object and surface with a static coefficient of friction $\mu_{\text{s}} = 0.100$. (a) How far can the spring be stretched without moving the mass? (b) If the object is set into oscillation with an amplitude twice the distance found in part (a), and the kinetic coefficient of friction is $\mu_{\text{k}} = 0.0850$, what total distance does it travel before stopping? Assume it starts at the maximum amplitude.

Solution:

a. 4.90×10^{-3} m; b. 1.15×10^{-2} m

Additional Problems

Exercise:

Problem:

Suppose you attach an object with mass *m* to a vertical spring originally at rest, and let it bounce up and down. You release the object from rest at the spring's original rest length, the length of the spring in equilibrium, without the mass attached. The amplitude of the motion is the distance between the equilibrium position of the spring without the mass attached and the equilibrium position of the spring with the mass attached. (a) Show that the spring exerts an upward force of 2.00*mg* on the object at its lowest point. (b) If the spring has a force constant of 10.0 N/m, is hung horizontally, and the position of the free end of the spring is marked as $y = 0.00\ \text{m}$, where is the new equilibrium position if a 0.25-kg-mass object is hung from the spring? (c) If the spring has a force constant of 10.0 M/m and a 0.25-kg-mass object is set in motion as described, find the amplitude of the oscillations. (d) Find the maximum velocity.

Exercise:

Problem:

A diver on a diving board is undergoing SHM. Her mass is 55.0 kg and the period of her motion is 0.800 s. The next diver is a male whose period of simple harmonic oscillation is 1.05 s. What is his mass if the mass of the board is negligible?

Solution:

94.7 kg

Exercise:

Problem:

Suppose a diving board with no one on it bounces up and down in a SHM with a frequency of 4.00 Hz. The board has an effective mass of 10.0 kg. What is the frequency of the SHM of a 75.0-kg diver on the board?

Exercise:

Problem:

The device pictured in the following figure entertains infants while keeping them from wandering. The child bounces in a harness suspended from a door frame by a spring. (a) If the spring stretches 0.250 m while supporting an 8.0-kg child, what is its force constant? (b) What is the time for one complete bounce of this child? (c) What is the child’s maximum velocity if the amplitude of her bounce is 0.200 m?

(credit: Lisa Doehnert)

Solution:

a. 314 N/m; b. 1.00 s; c. 1.25 m/s

Exercise:

Problem:

A mass is placed on a frictionless, horizontal table. A spring $(k = 100\,\mathrm{N/m})$, which can be stretched or compressed, is placed on the table. A 5.00-kg mass is attached to one end of the spring, the other end is anchored to the wall. The equilibrium position is marked at zero. A student moves the mass out to $x = 4.0\mathrm{cm}$ and releases it from rest. The mass oscillates in SHM. (a) Determine the equations of motion. (b) Find the position, velocity, and acceleration of the mass at time $t = 3.00\,\mathrm{s}$.

Exercise:

Problem:

Find the ratio of the new/old periods of a pendulum if the pendulum were transported from Earth to the Moon, where the acceleration due to gravity is $1.63\,\mathrm{m/s}^2$.

Solution:

ratio of 2.45

Exercise:

Problem:

At what rate will a pendulum clock run on the Moon, where the acceleration due to gravity is $1.63\,\mathrm{m/s}^2$, if it keeps time accurately on Earth? That is, find the time (in hours) it takes the clock’s hour hand to make one revolution on the Moon.

Exercise:

Problem:

If a pendulum-driven clock gains 5.00 s/day, what fractional change in pendulum length must be made for it to keep perfect time?

Solution:

The length must increase by 0.0116%.

Exercise:

Problem:

A 2.00-kg object hangs, at rest, on a 1.00-m-long string attached to the ceiling. A 100-g mass is fired with a speed of 20 m/s at the 2.00-kg mass, and the 100.00-g mass collides perfectly elastically with the 2.00-kg mass. Write an equation for the motion of the hanging mass after the collision. Assume air resistance is negligible.

Exercise:

Problem:

A 2.00-kg object hangs, at rest, on a 1.00-m-long string attached to the ceiling. A 100-g object is fired with a speed of 20 m/s at the 2.00-kg object, and the two objects collide and stick together in a totally inelastic collision. Write an equation for the motion of the system after the collision. Assume air resistance is negligible.

Solution:

$\theta = (0.31\ \text{rad})\text{sin}\left(3.13\ \text{s}^{-1}t\right)$

Exercise:

Problem:

Assume that a pendulum used to drive a grandfather clock has a length $L_0 = 1.00\ \text{m}$ and a mass M at temperature $T = 20.00^\circ\text{C}$. It can be modeled as a physical pendulum as a rod oscillating around one end. By what percentage will the period change if the temperature increases by 10°C? Assume the length of the rod changes linearly with temperature, where $L = L_0\left(1 + \alpha\Delta T\right)$ and the rod is made of brass $\left(\alpha = 18\ \times\ 10^{-6\,\circ}\text{C}^{-1}\right)$.

Exercise:

Problem:

A 2.00-kg block lies at rest on a frictionless table. A spring, with a spring constant of 100 N/m is attached to the wall and to the block. A second block of 0.50 kg is placed on top of the first block. The 2.00-kg block is gently pulled to a position $x = +A$ and released from rest. There is a coefficient of friction of 0.45 between the two blocks. (a) What is the period of the oscillations? (b) What is the largest amplitude of motion that will allow the blocks to oscillate without the 0.50-kg block sliding off?

Solution:

a. 0.99 s; b. 0.11 m

Challenge Problems

Exercise:

Problem:

A suspension bridge oscillates with an effective force constant of $1.00 \times 10^8 \text{ N/m}$. (a) How much energy is needed to make it oscillate with an amplitude of 0.100 m? (b) If soldiers march across the bridge with a cadence equal to the bridge's natural frequency and impart 1.00×10^4 J of energy each second, how long does it take for the bridge's oscillations to go from 0.100 m to 0.500 m amplitude.

Exercise:

Problem:

Near the top of the Citigroup Center building in New York City, there is an object with mass of 4.00×10^5 kg on springs that have adjustable force constants. Its function is to dampen wind-driven oscillations of the building by oscillating at the same frequency as the building is being driven—the driving force is transferred to the object, which oscillates instead of the entire building. (a) What effective force constant should the springs have to make the object oscillate with a period of 2.00 s? (b) What energy is stored in the springs for a 2.00-m displacement from equilibrium?

Solution:

a. 3.95×10^6 N/m; b. 7.90×10^6 J

Exercise:

Problem:

Parcels of air (small volumes of air) in a stable atmosphere (where the temperature increases with height) can oscillate up and down, due to the restoring force provided by the buoyancy of the air parcel. The frequency of the oscillations are a measure of the stability of the atmosphere. Assuming that the acceleration of an air parcel can be modeled as $\frac{\partial^2 z'}{\partial t^2} = \frac{g}{\rho_o} \frac{\partial \rho(z)}{\partial z} z'$, prove that $z' = z_0{}' e^{t\sqrt{-N^2}}$ is a solution, where *N* is known as the Brunt-Väisälä frequency. Note that in a stable atmosphere, the density decreases with height and parcel oscillates up and down.

Exercise:

Problem:

Consider the van der Waals potential $U(r) = U_o \left[\left(\frac{R_o}{r} \right)^{12} - 2 \left(\frac{R_o}{r} \right)^{6} \right]$, used to model the potential energy function of two molecules, where the minimum potential is at $r = R_o$. Find the force as a function of *r*. Consider a small displacement $r = R_o + r'$ and use the binomial theorem:

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots,$$

to show that the force does approximate a Hooke’s law force.

Solution:

$F \approx -\text{constant}\, r'$

Exercise:

Problem:

Suppose the length of a clock’s pendulum is changed by 1.000%, exactly at noon one day. What time will the clock read 24.00 hours later, assuming it the pendulum has kept perfect time before the change? Note that there are two answers, and perform the calculation to four-digit precision.

Exercise:

Problem:

(a) The springs of a pickup truck act like a single spring with a force constant of $1.30 \times 10^5\ \text{N/m}$. By how much will the truck be depressed by its maximum load of 1000 kg? (b) If the pickup truck has four identical springs, what is the force constant of each?

Solution:

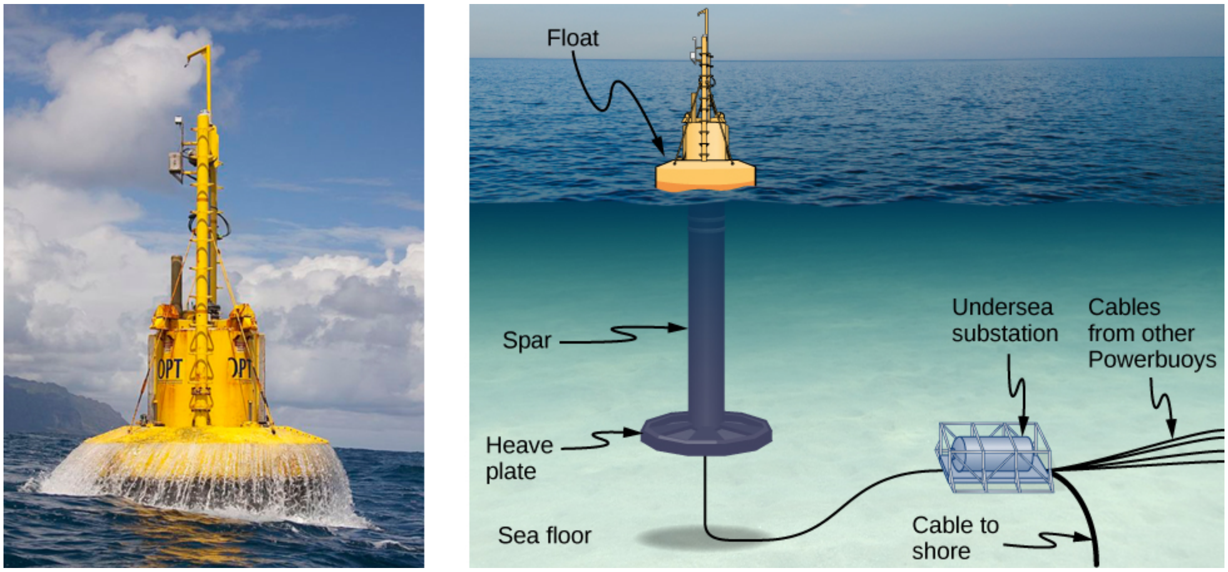
a. 7.54 cm; b. $3.25 \times 10^4\ \text{N/m}$

Glossary

resonance
large amplitude oscillations in a system produced by a small amplitude driving force, which has a frequency equal to the natural frequency

Introduction class="introduction"

From the world of renewable energy sources comes the electric power-generating buoy. Although there are many versions, this one converts the up-and-down motion, as well as side-to-side motion, of the buoy into rotational motion in order to turn an electric generator, which stores the energy in batteries.



In this chapter, we study the physics of wave motion. We concentrate on mechanical waves, which are disturbances that move through a medium such as air or water. Like simple harmonic motion studied in the preceding chapter, the energy transferred through the medium is proportional to the amplitude squared. Surface water waves in the ocean are transverse waves in which the energy of the wave travels horizontally while the water oscillates up and down due to some restoring force. In the picture above, a buoy is used to convert the awesome power of ocean waves into electricity. The up-and-down motion of the buoy generated as the waves pass is converted into rotational motion that turns a rotor in an electric generator. The generator charges batteries, which are in turn used to provide a consistent energy source for the end user. This model was successfully tested by the US Navy in a project to provide power to coastal security networks and was able to provide an average power of 350 W. The buoy survived the difficult ocean environment, including operation off the New Jersey coast through Hurricane Irene in 2011.

The concepts presented in this chapter will be the foundation for many interesting topics, from the transmission of information to the concepts of quantum mechanics.

Traveling Waves

By the end of this section, you will be able to:

- Describe the basic characteristics of wave motion
- Define the terms wavelength, amplitude, period, frequency, and wave speed
- Explain the difference between longitudinal and transverse waves, and give examples of each type
- List the different types of waves

We saw in Oscillations that oscillatory motion is an important type of behavior that can be used to model a wide range of physical phenomena. Oscillatory motion is also important because oscillations can generate waves, which are of fundamental importance in physics. Many of the terms and equations we studied in the chapter on oscillations apply equally well to wave motion ([link]).

An ocean wave is probably the first picture that comes to mind when you hear the word “wave.” Although this breaking wave, and ocean waves in general, have apparent similarities to the basic wave characteristics we will discuss, the mechanisms driving ocean waves are highly complex and beyond the scope of this chapter. It may seem natural, and even advantageous, to apply the concepts in this chapter to ocean waves, but ocean waves are nonlinear, and the simple models presented in this chapter do not fully explain them. (credit: Steve Jurvetson)

Types of Waves

A **wave** is a disturbance that propagates, or moves from the place it was created. There are three basic types of waves: mechanical waves, electromagnetic waves, and matter waves.

Basic **mechanical waves** are governed by Newton’s laws and require a medium. A medium is the substance a mechanical waves propagates through, and the medium produces an elastic restoring force when it is deformed. Mechanical waves transfer energy and momentum, without transferring mass. Some examples of mechanical waves are water waves, sound waves, and seismic waves. The medium for water waves is water; for sound waves, the medium is usually air. (Sound waves can travel in other media as well; we will look at that in more detail in Sound.) For surface water waves, the disturbance occurs on the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker or a vibrating tuning fork. In both cases, the disturbance is the oscillation of the molecules of the fluid. In mechanical waves, energy and momentum transfer with the motion of the wave, whereas the mass oscillates around an equilibrium point. (We discuss this in Energy and Power of a Wave.) Earthquakes generate seismic waves from several types of disturbances, including the disturbance of Earth’s surface and pressure disturbances under the surface. Seismic waves travel through the solids and liquids that form Earth. In this chapter, we focus on mechanical waves.

Electromagnetic waves are associated with oscillations in electric and magnetic fields and do not require a medium. Examples include gamma rays, X-rays, ultraviolet waves, visible light, infrared waves, microwaves, and radio waves. Electromagnetic waves can travel through a vacuum at the speed of light, $v = c = 2.99792458\ \times\ 10^8\ \text{m/s}$. For example, light from distant stars travels through the vacuum of space and reaches Earth. Electromagnetic waves have some characteristics that are similar to mechanical waves; they are covered in more detail in Electromagnetic Waves.

Matter waves are a central part of the branch of physics known as quantum mechanics. These waves are associated with protons, electrons, neutrons, and other fundamental particles found in nature. The theory that all types of matter have wave-like properties was first proposed by Louis de Broglie in 1924. Matter waves are discussed in Photons and Matter Waves.

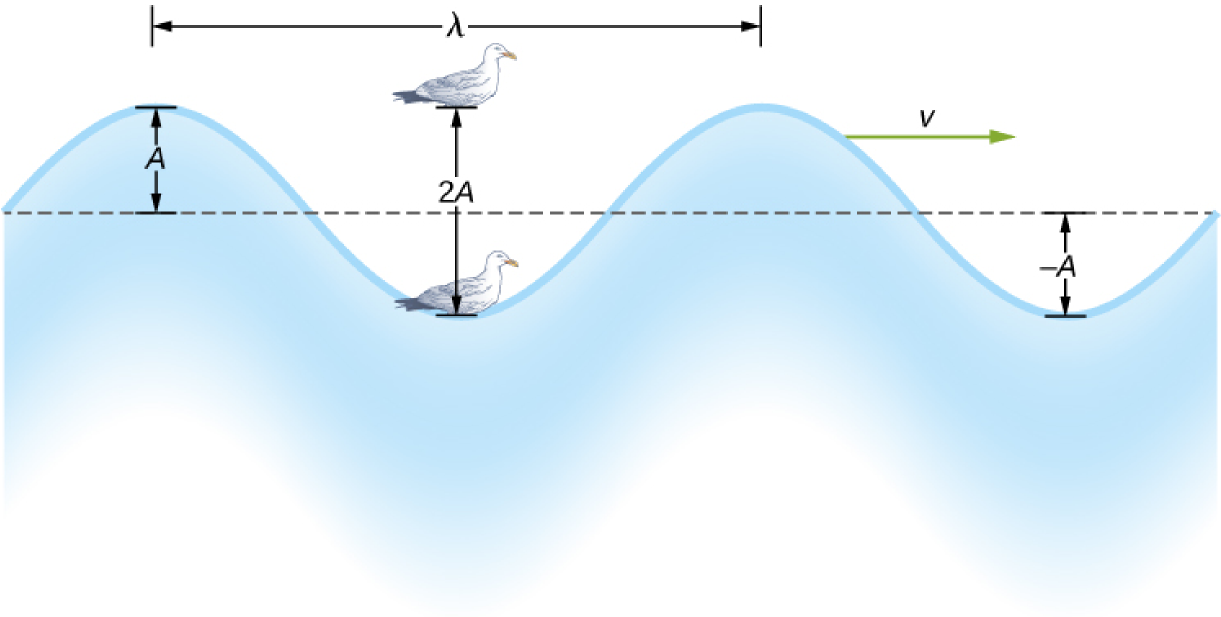
Mechanical Waves

Mechanical waves exhibit characteristics common to all waves, such as amplitude, wavelength, period, frequency, and energy. All wave characteristics can be described by a small set of underlying principles.

The simplest mechanical waves repeat themselves for several cycles and are associated with simple harmonic motion. These simple harmonic waves can be modeled using some combination of sine and cosine functions. For example, consider the simplified surface water wave that moves across the surface of water as illustrated in [link]. Unlike complex ocean waves, in surface water waves, the medium, in this case water, moves vertically, oscillating up and down, whereas the disturbance of the wave moves horizontally through the medium. In [link], the waves causes a seagull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The crest is the highest point of the wave, and the trough is the lowest part of the wave. The time for one complete oscillation of the up-and-down motion is the wave’s period T. The wave’s frequency is the number of waves that pass through a point per unit time and is equal to $f = 1/T$. The period can be expressed using any convenient unit of time but is usually measured in seconds; frequency is usually measured in hertz (Hz), where $1\text{ Hz}=1\text{ s}^{-1}$.

The length of the wave is called the **wavelength** and is represented by the Greek letter lambda (λ), which is measured in any convenient unit of length, such as a centimeter or meter. The wavelength can be measured between any two similar points along the medium that have the same height and the same slope. In [link], the wavelength is shown measured between two crests. As stated above, the period of the wave is equal to the time for one oscillation, but it is also equal to the time for one wavelength to pass through a point along the wave's path.

The amplitude of the wave (A) is a measure of the maximum displacement of the medium from its equilibrium position. In the figure, the equilibrium position is indicated by the dotted line, which is the height of the water if there were no waves moving through it. In this case, the wave is symmetrical, the crest of the wave is a distance $+A$ above the equilibrium position, and the trough is a distance $-A$ below the equilibrium position. The units for the amplitude can be centimeters or meters, or any convenient unit of distance.



An idealized surface water wave passes under a seagull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The amplitude A of the wave is the maximum displacement of the wave from the equilibrium position, which is indicated by the dotted line. In this example, the medium moves up and down, whereas the disturbance of the surface propagates parallel to the surface at a speed v.

The water wave in the figure moves through the medium with a propagation velocity $\vec{\mathbf{v}}$. The magnitude of the **wave velocity** is the distance the wave travels in a given time, which is one wavelength in the time of one period, and the **wave speed** is the magnitude of wave velocity. In equation form, this is

Note:
Equation:

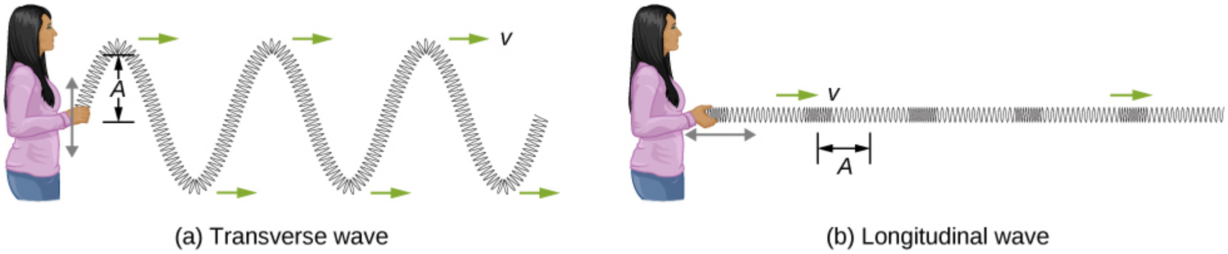
$$v = \frac{\lambda}{T} = \lambda f.$$

This fundamental relationship holds for all types of waves. For water waves, v is the speed of a surface wave; for sound, v is the speed of sound; and for visible light, v is the speed of light.

Transverse and Longitudinal Waves

We have seen that a simple mechanical wave consists of a periodic disturbance that propagates from one place to another through a medium. In [link](a), the wave propagates in the horizontal direction, whereas the medium is disturbed in the vertical direction. Such a wave is called a

transverse wave. In a transverse wave, the wave may propagate in any direction, but the disturbance of the medium is perpendicular to the direction of propagation. In contrast, in a **longitudinal wave** or compressional wave, the disturbance is parallel to the direction of propagation. [link](b) shows an example of a longitudinal wave. The size of the disturbance is its amplitude *A* and is completely independent of the speed of propagation *v*.

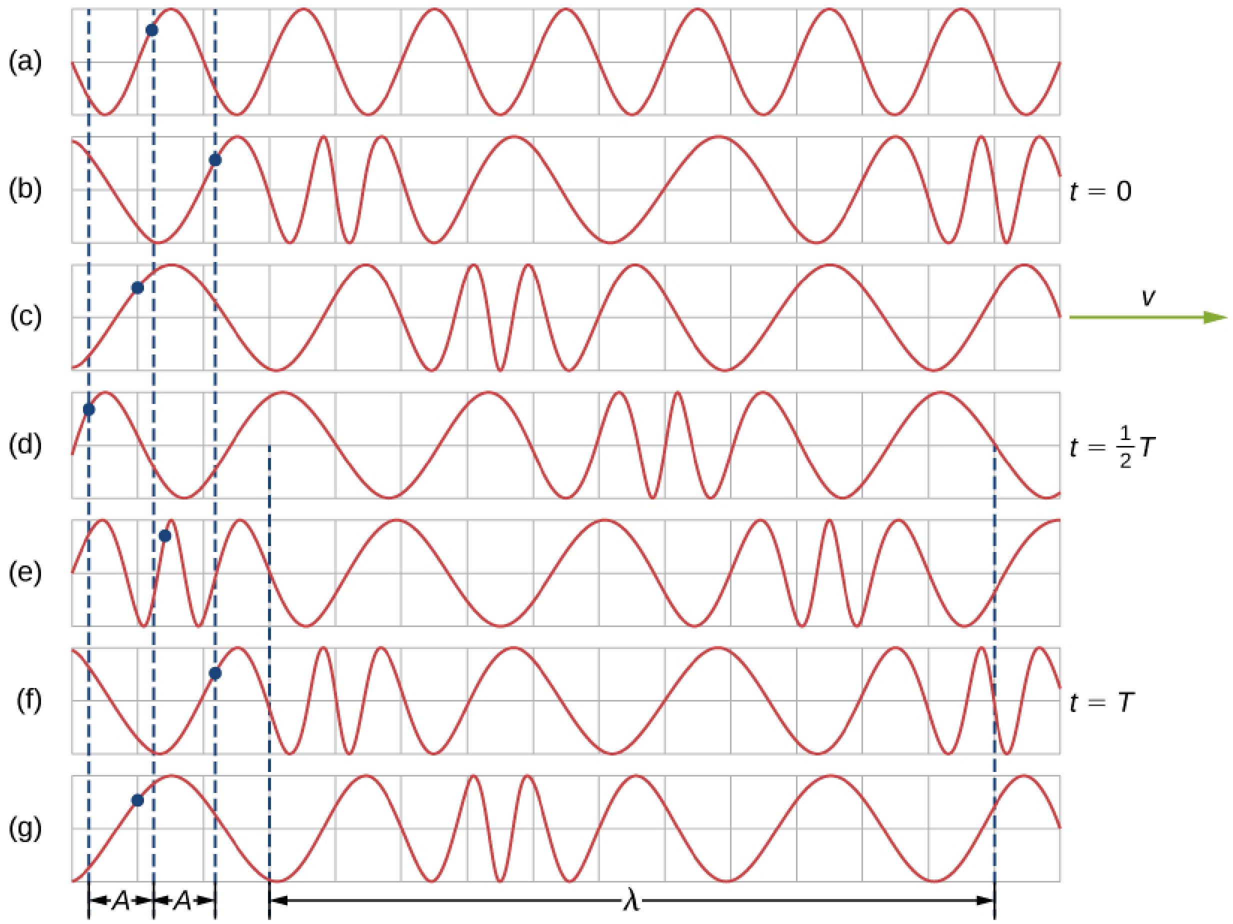


(a) Transverse wave

(b) Longitudinal wave

(a) In a transverse wave, the medium oscillates perpendicular to the wave velocity. Here, the spring moves vertically up and down, while the wave propagates horizontally to the right. (b) In a longitudinal wave, the medium oscillates parallel to the propagation of the wave. In this case, the spring oscillates back and forth, while the wave propagates to the right.

A simple graphical representation of a section of the spring shown in [link](b) is shown in [link]. [link](a) shows the equilibrium position of the spring before any waves move down it. A point on the spring is marked with a blue dot. [link](b) through (g) show snapshots of the spring taken one-quarter of a period apart, sometime after the end of` the spring is oscillated back and forth in the *x*-direction at a constant frequency. The disturbance of the wave is seen as the compressions and the expansions of the spring. Note that the blue dot oscillates around its equilibrium position a distance *A*, as the longitudinal wave moves in the positive *x*-direction with a constant speed. The distance *A* is the amplitude of the wave. The *y*-position of the dot does not change as the wave moves through the spring. The wavelength of the wave is measured in part (d). The wavelength depends on the speed of the wave and the frequency of the driving force.



(a) This is a simple, graphical representation of a section of the stretched spring shown in [link](b), representing the spring’s equilibrium position before any waves are induced on the spring. A point on the spring is marked by a blue dot. (b–g) Longitudinal waves are created by oscillating the end of the spring (not shown) back and forth along the *x*-axis. The longitudinal wave, with a wavelength λ, moves along the spring in the +*x*-direction with a wave speed *v*. For convenience, the wavelength is measured in (d). Note that the point on the spring that was marked with the blue dot moves back and forth a distance *A* from the equilibrium position, oscillating around the equilibrium position of the point.

Waves may be transverse, longitudinal, or a combination of the two. Examples of transverse waves are the waves on stringed instruments or surface waves on water, such as ripples moving on a pond. Sound waves in air and water are longitudinal. With sound waves, the disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and for this reason, the sound waves in them are longitudinal waves. Sound in solids can have both longitudinal and transverse components, such as those in a seismic wave. Earthquakes generate seismic waves under Earth's surface with both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). The components of seismic waves have important individual characteristics—they propagate at different speeds, for example. Earthquakes also have surface waves that are similar to surface waves on water. Ocean waves also have both transverse and longitudinal components.

Example:

Wave on a String

A student takes a 30.00-m-long string and attaches one end to the wall in the physics lab. The student then holds the free end of the rope, keeping the tension constant in the rope. The student then begins to send waves down the string by moving the end of the string up and down with a frequency of 2.00 Hz. The maximum displacement of the end of the string is 20.00 cm. The first wave hits the lab wall 6.00 s after it was created. (a) What is the speed of the wave? (b) What is the period of the wave? (c) What is the wavelength of the wave?

Strategy

a. The speed of the wave can be derived by dividing the distance traveled by the time.
b. The period of the wave is the inverse of the frequency of the driving force.
c. The wavelength can be found from the speed and the period $v = \lambda / T$.

Solution

a. The first wave traveled 30.00 m in 6.00 s:
Equation:

$$v=\frac{30.00\,\text{m}}{6.00\,\text{s}}=5.00\frac{\text{m}}{\text{s}}.$$

b. The period is equal to the inverse of the frequency:
Equation:

$$T=\frac{1}{f}=\frac{1}{2.00\,\text{s}^{-1}}=0.50\,\text{s}.$$

c. The wavelength is equal to the velocity times the period:
Equation:

$$\lambda=vT=5.00\frac{\text{m}}{\text{s}}(0.50\,\text{s})=2.50\,\text{m}.$$

Significance
The frequency of the wave produced by an oscillating driving force is equal to the frequency of the driving force.

Note:
Exercise:

Problem:

Check Your Understanding When a guitar string is plucked, the guitar string oscillates as a result of waves moving through the string. The vibrations of the string cause the air molecules to oscillate, forming sound waves. The frequency of the sound waves is equal to the frequency of the vibrating string. Is the wavelength of the sound wave always equal to the wavelength of the waves on the string?

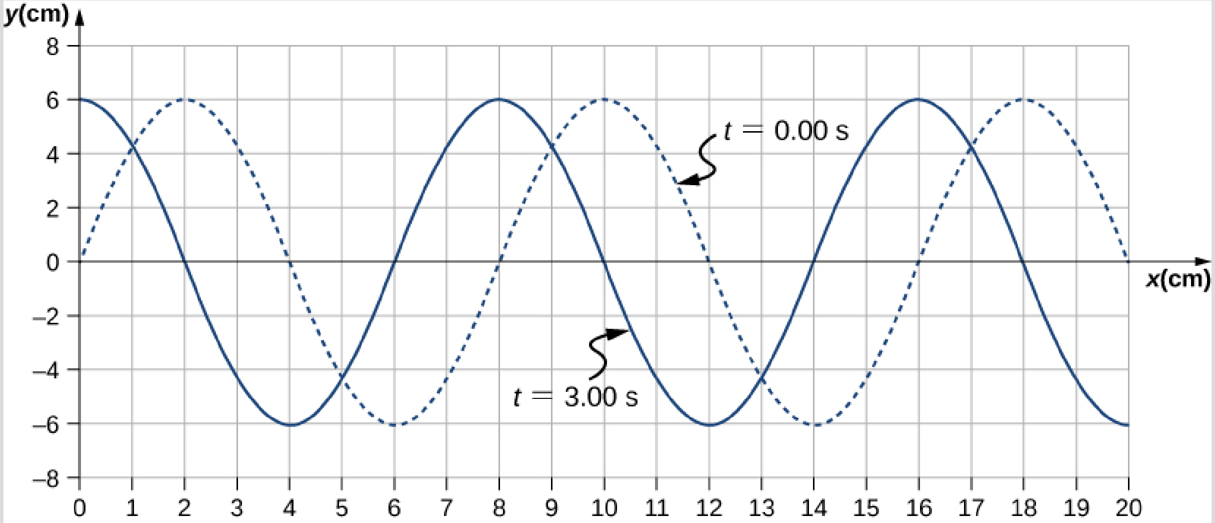
Solution:

The wavelength of the waves depends on the frequency and the velocity of the wave. The frequency of the sound wave is equal to the frequency of the wave on the string. The wavelengths of the sound waves and the waves on the string are equal only if the velocities of the waves are the same, which is not always the case. If the speed of the sound wave is different from the speed of the wave on the string, the wavelengths are different. This velocity of sound waves will be discussed in Sound.

Example:

Characteristics of a Wave

A transverse mechanical wave propagates in the positive *x*-direction through a spring (as shown in [link](a)) with a constant wave speed, and the medium oscillates between $+A$ and $-A$ around an equilibrium position. The graph in [link] shows the height of the spring (*y*) versus the position (*x*), where the *x*-axis points in the direction of propagation. The figure shows the height of the spring versus the *x*-position at $t = 0.00$ s as a dotted line and the wave at $t = 3.00$ s as a solid line. Assume the wave has not traveled more than 1 wavelength in this time. (a) Determine the wavelength and amplitude of the wave. (b) Find the propagation velocity of the wave. (c) Calculate the period and frequency of the wave.



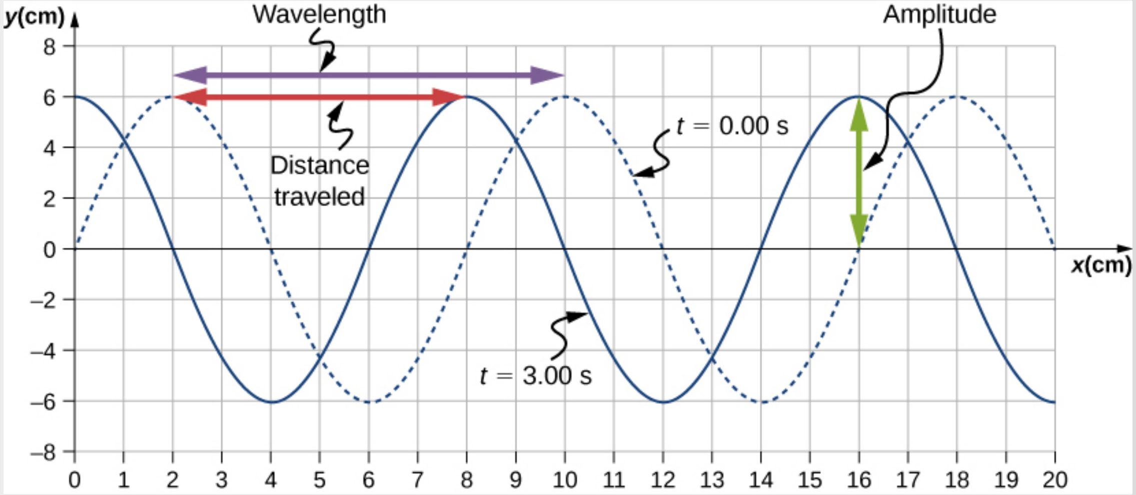
A transverse wave shown at two instants of time.

Strategy

a. The amplitude and wavelength can be determined from the graph.
b. Since the velocity is constant, the velocity of the wave can be found by dividing the distance traveled by the wave by the time it took the wave to travel the distance.
c. The period can be found from $v = \frac{\lambda}{T}$ and the frequency from $f = \frac{1}{T}$.

Solution

a. Read the wavelength from the graph, looking at the purple arrow in [link]. Read the amplitude by looking at the green arrow. The wavelength is $\lambda = 8.00 \text{ cm}$ and the amplitude is $A = 6.00 \text{ cm}$.



Characteristics of the wave marked on a graph of its displacement.

b. The distance the wave traveled from time $t = 0.00 \text{ s}$ to time $t = 3.00 \text{ s}$ can be seen in the graph. Consider the red arrow, which shows the distance the crest has moved in 3 s. The distance is $8.00 \text{ cm} - 2.00 \text{ cm} = 6.00 \text{ cm}$. The velocity is

Equation:

$$v = \frac{\Delta x}{\Delta t} = \frac{8.00\text{ cm} - 2.00\text{ cm}}{3.00\text{ s} - 0.00\text{ s}} = 2.00\text{ cm/s}.$$

c. The period is $T = \frac{\lambda}{v} = \frac{8.00\text{ cm}}{2.00\text{ cm/s}} = 4.00\text{ s}$ and the frequency is $f = \frac{1}{T} = \frac{1}{4.00\text{ s}} = 0.25\text{ Hz}.$

Significance
Note that the wavelength can be found using any two successive identical points that repeat, having the same height and slope. You should choose two points that are most convenient. The displacement can also be found using any convenient point.

Note:
Exercise:

Problem:

Check Your Understanding The propagation velocity of a transverse or longitudinal mechanical wave may be constant as the wave disturbance moves through the medium. Consider a transverse mechanical wave: Is the velocity of the medium also constant?

Solution:

In a transverse wave, the wave may move at a constant propagation velocity through the medium, but the medium oscillates perpendicular to the motion of the wave. If the wave moves in the positive *x*-direction, the medium oscillates up and down in the *y*-direction. The velocity of the medium is therefore not constant, but the medium’s velocity and acceleration are similar to that of the simple harmonic motion of a mass on a spring.

Summary

- A wave is a disturbance that moves from the point of origin with a wave velocity v.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. Wave velocity and wavelength are related to the wave's frequency and period by $v = \frac{\lambda}{T} = \lambda f$.
- Mechanical waves are disturbances that move through a medium and are governed by Newton's laws.
- Electromagnetic waves are disturbances in the electric and magnetic fields, and do not require a medium.
- Matter waves are a central part of quantum mechanics and are associated with protons, electrons, neutrons, and other fundamental particles found in nature.
- A transverse wave has a disturbance perpendicular to the wave's direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Conceptual Questions

Exercise:

Problem:

Give one example of a transverse wave and one example of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.

Solution:

A wave on a guitar string is an example of a transverse wave. The disturbance of the string moves perpendicular to the propagation of the wave. The sound produced by the string is a longitudinal wave where the disturbance of the air moves parallel to the propagation of the wave.

Exercise:

Problem:

A sinusoidal transverse wave has a wavelength of 2.80 m. It takes 0.10 s for a portion of the string at a position *x* to move from a maximum position of $y = 0.03\ \mathrm{m}$ to the equilibrium position $y = 0$. What are the period, frequency, and wave speed of the wave?

Exercise:

Problem:

What is the difference between propagation speed and the frequency of a mechanical wave? Does one or both affect wavelength? If so, how?

Solution:

Propagation speed is the speed of the wave propagating through the medium. If the wave speed is constant, the speed can be found by $v = \frac{\lambda}{T} = \lambda f$. The frequency is the number of wave that pass a point per unit time. The wavelength is directly proportional to the wave speed and inversely proportional to the frequency.

Exercise:

Problem:

Consider a stretched spring, such as a slinky. The stretched spring can support longitudinal waves and transverse waves. How can you produce transverse waves on the spring? How can you produce longitudinal waves on the spring?

Exercise:

Problem:

Consider a wave produced on a stretched spring by holding one end and shaking it up and down. Does the wavelength depend on the distance you move your hand up and down?

Solution:

No, the distance you move your hand up and down will determine the amplitude of the wave. The wavelength will depend on the frequency you move your hand up and down, and the speed of the wave through the spring.

Exercise:

Problem:

A sinusoidal, transverse wave is produced on a stretched spring, having a period *T*. Each section of the spring moves perpendicular to the direction of propagation of the wave, in simple harmonic motion with an amplitude *A*. Does each section oscillate with the same period as the wave or a different period? If the amplitude of the transverse wave were doubled but the period stays the same, would your answer be the same?

Exercise:

Problem:

An electromagnetic wave, such as light, does not require a medium. Can you think of an example that would support this claim?

Solution:

Light from the Sun and stars reach Earth through empty space where there is no medium present.

Problems

Exercise:

Problem:

Storms in the South Pacific can create waves that travel all the way to the California coast, 12,000 km away. How long does it take them to travel this distance if they travel at 15.0 m/s?

Exercise:

Problem:

Waves on a swimming pool propagate at 0.75 m/s. You splash the water at one end of the pool and observe the wave go to the opposite end, reflect, and return in 30.00 s. How far away is the other end of the pool?

Solution:

$2d = vt \Rightarrow d = 11.25\,\mathrm{m}$

Exercise:

Problem:

Wind gusts create ripples on the ocean that have a wavelength of 5.00 cm and propagate at 2.00 m/s. What is their frequency?

Exercise:

Problem:

How many times a minute does a boat bob up and down on ocean waves that have a wavelength of 40.0 m and a propagation speed of 5.00 m/s?

Solution:

$$\begin{aligned} v &= f\lambda,\text{ so that } f = 0.125\,\mathrm{Hz},\text{ so that} \\ N &= 7.50\,\text{times} \end{aligned}$$

Exercise:

Problem:

Scouts at a camp shake the rope bridge they have just crossed and observe the wave crests to be 8.00 m apart. If they shake the bridge twice per second, what is the propagation speed of the waves?

Exercise:

Problem:

What is the wavelength of the waves you create in a swimming pool if you splash your hand at a rate of 2.00 Hz and the waves propagate at a wave speed of 0.800 m/s?

Solution:

$v=f\lambda \Rightarrow \lambda = 0.400\,\text{m}$

Exercise:

Problem:

What is the wavelength of an earthquake that shakes you with a frequency of 10.0 Hz and gets to another city 84.0 km away in 12.0 s?

Exercise:

Problem:

Radio waves transmitted through empty space at the speed of light $\left(v=c=3.00\ \times\ 10^{8}\ \text{m/s}\right)$ by the *Voyager* spacecraft have a wavelength of 0.120 m. What is their frequency?

Solution:

$v=f\lambda \Rightarrow f=2.50\ \times\ 10^{9}\ \text{Hz}$

Exercise:

Problem:

Your ear is capable of differentiating sounds that arrive at each ear just 0.34 ms apart, which is useful in determining where low frequency sound is originating from. (a) Suppose a low-frequency sound source is placed to the right of a person, whose ears are approximately 18 cm apart, and the speed of sound generated is 340 m/s. How long is the interval between when the sound arrives at the right ear and the sound arrives at the left ear? (b) Assume the same person was scuba diving and a low-frequency sound source was to the right of the scuba diver. How long is the interval between when the sound arrives at the right ear and the sound arrives at the left ear, if the speed of sound in water is 1500 m/s? (c) What is significant about the time interval of the two situations?

Exercise:

Problem:

(a) Seismographs measure the arrival times of earthquakes with a precision of 0.100 s. To get the distance to the epicenter of the quake, geologists compare the arrival times of S- and P-waves, which travel at different speeds. If S- and P-waves travel at 4.00 and 7.20 km/s, respectively, in the region considered, how precisely can the distance to the source of the earthquake be determined? (b) Seismic waves from underground detonations of nuclear bombs can be used to locate the test site and detect violations of test bans. Discuss whether your answer to (a) implies a serious limit to such detection. (Note also that the uncertainty is greater if there is an uncertainty in the propagation speeds of the S- and P-waves.)

Solution:

a. The P-waves outrun the S-waves by a speed of $v = 3.20\,\text{km/s}$; therefore, $\Delta d = 0.320\,\text{km}$. b. Since the uncertainty in the distance is less than a kilometer, our answer to part (a) does not seem to limit the detection of nuclear bomb detonations. However, if the velocities are uncertain, then the uncertainty in the distance would increase and could then make it difficult to identify the source of the seismic waves.

Exercise:

Problem:

A Girl Scout is taking a 10.00-km hike to earn a merit badge. While on the hike, she sees a cliff some distance away. She wishes to estimate the time required to walk to the cliff. She knows that the speed of sound is approximately 343 meters per second. She yells and finds that the echo returns after approximately 2.00 seconds. If she can hike 1.00 km in 10 minutes, how long would it take her to reach the cliff?

Exercise:

Problem:

A quality assurance engineer at a frying pan company is asked to qualify a new line of nonstick-coated frying pans. The coating needs to be 1.00 mm thick. One method to test the thickness is for the engineer to pick a percentage of the pans manufactured, strip off the coating, and measure the thickness using a micrometer. This method is a destructive testing method. Instead, the engineer decides that every frying pan will be tested using a nondestructive method. An ultrasonic transducer is used that produces sound waves with a frequency of $f = 25\,\text{kHz}$. The sound waves are sent through the coating and are reflected by the interface between the coating and the metal pan, and the time is recorded. The wavelength of the ultrasonic waves in the coating is 0.076 m. What should be the time recorded if the coating is the correct thickness (1.00 mm)?

Solution:

$$\begin{aligned} v &= 1900\,\text{m/s} \\ \Delta t &= 1.05\,\mu\,\text{s} \end{aligned}$$

Glossary

longitudinal wave
wave in which the disturbance is parallel to the direction of propagation

mechanical wave
wave that is governed by Newton’s laws and requires a medium

transverse wave
wave in which the disturbance is perpendicular to the direction of propagation

wave
disturbance that moves from its source and carries energy

wave velocity
velocity at which the disturbance moves; also called the propagation velocity

wave speed
magnitude of the wave velocity

wavelength
distance between adjacent identical parts of a wave

Mathematics of Waves

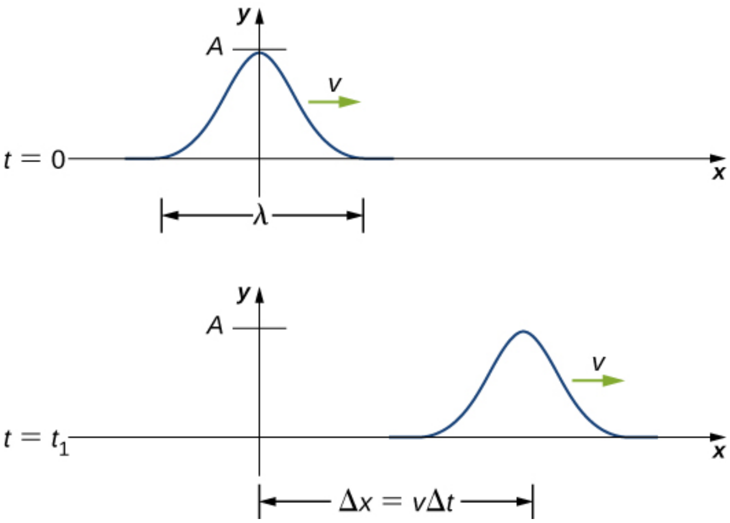
By the end of this section, you will be able to:

- Model a wave, moving with a constant wave velocity, with a mathematical expression
- Calculate the velocity and acceleration of the medium
- Show how the velocity of the medium differs from the wave velocity (propagation velocity)

In the previous section, we described periodic waves by their characteristics of wavelength, period, amplitude, and wave speed of the wave. Waves can also be described by the motion of the particles of the medium through which the waves move. The position of particles of the medium can be mathematically modeled as **wave functions**, which can be used to find the position, velocity, and acceleration of the particles of the medium of the wave at any time.

Pulses

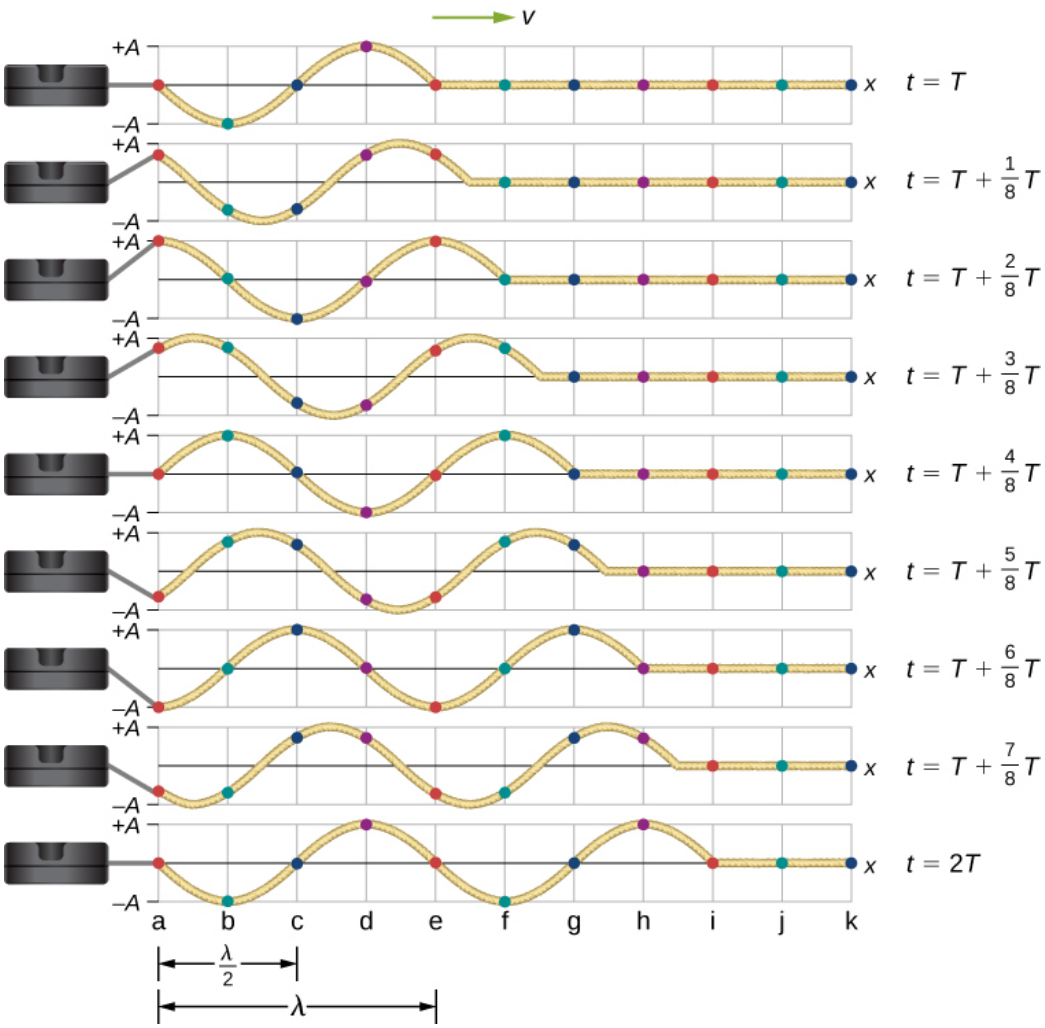
A **pulse** can be described as wave consisting of a single disturbance that moves through the medium with a constant amplitude. The pulse moves as a pattern that maintains its shape as it propagates with a constant wave speed. Because the wave speed is constant, the distance the pulse moves in a time Δt is equal to $\Delta x = v\Delta t$ ([link]).



The pulse at time $t = 0$ is centered on $x = 0$ with amplitude A. The pulse moves as a pattern with a constant shape, with a constant maximum value A. The velocity is constant and the pulse moves a distance $\Delta x = v\Delta t$ in a time Δt. The distance traveled is measured with any convenient point on the pulse. In this figure, the crest is used.

Modeling a One-Dimensional Sinusoidal Wave using a Wave Function

Consider a string kept at a constant tension F_T where one end is fixed and the free end is oscillated between $y = +A$ and $y = -A$ by a mechanical device at a constant frequency. [link] shows snapshots of the wave at an interval of an eighth of a period, beginning after one period $(t = T)$.



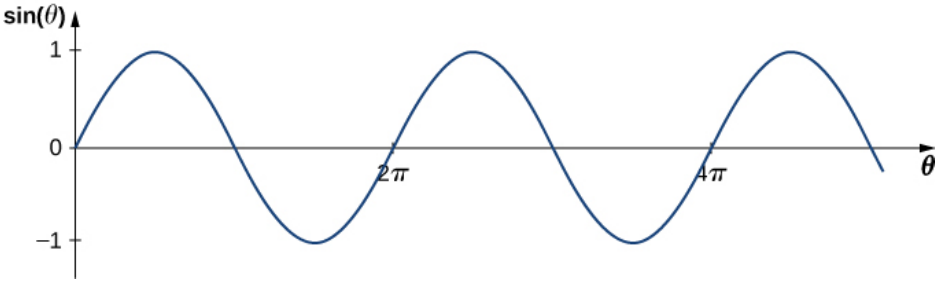
Snapshots of a transverse wave moving through a string under tension, beginning at time $t = T$ and taken at intervals of $\frac{1}{8}T$. Colored dots are used to highlight points on the string. Points that are a wavelength apart in the *x*-direction are highlighted with the same color dots.

Notice that each select point on the string (marked by colored dots) oscillates up and down in simple harmonic motion, between $y = +A$ and $y = -A$, with a period T. The wave on the string is sinusoidal and is translating in the positive *x*-direction as time progresses.

At this point, it is useful to recall from your study of algebra that if *f*(*x*) is some function, then $f\left(x - d\right)$ is the same function translated in the positive *x*-direction by a distance *d*. The function $f\left(x + d\right)$ is the same function translated in the negative *x*-direction by a distance *d*. We want to define a wave function that will give the *y*-position of each segment of the string for every position *x* along the string for every time *t*.

Looking at the first snapshot in [link], the *y*-position of the string between $x = 0$ and $x = \lambda$ can be modeled as a sine function. This wave propagates down the string one wavelength in one period, as seen in the last snapshot. The wave therefore moves with a constant wave speed of $v = \lambda / T$.

Recall that a sine function is a function of the angle θ, oscillating between $+1$ and -1, and repeating every 2π radians ([link]). However, the *y*-position of the medium, or the wave function, oscillates between $+A$ and $-A$, and repeats every wavelength λ.



A sine function oscillates between $+1$ and -1 every 2π radians.

To construct our model of the wave using a periodic function, consider the ratio of the angle and the position,

Equation:

$$\begin{aligned}\frac{\theta}{x} &= \frac{2\pi}{\lambda},\\ \theta &= \frac{2\pi}{\lambda}x.\end{aligned}$$

Using $\theta = \frac{2\pi}{\lambda}x$ and multiplying the sine function by the amplitude *A*, we can now model the *y*-position of the string as a function of the position *x*:

Equation:

$$y\left(x\right) = A\,\text{sin}\left(\frac{2\pi}{\lambda}x\right).$$

The wave on the string travels in the positive *x*-direction with a constant velocity *v*, and moves a distance *vt* in a time *t*. The wave function can now be defined by

Equation:

$$y\left(x,t\right) = A\,\text{sin}\left(\frac{2\pi}{\lambda}\left(x - vt\right)\right).$$

It is often convenient to rewrite this wave function in a more compact form. Multiplying through by the ratio $\frac{2\pi}{\lambda}$ leads to the equation

Equation:

$$y\left(x,t\right) = A\,\text{sin}\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{\lambda}vt\right).$$

The value $\frac{2\pi}{\lambda}$ is defined as the **wave number**. The symbol for the wave number is *k* and has units of inverse meters, m^{-1}:

Note:
Equation:

$$k \equiv \frac{2\pi}{\lambda}$$

Recall from Oscillations that the angular frequency is defined as $\omega \equiv \frac{2\pi}{T}$. The second term of the wave function becomes
Equation:

$$\frac{2\pi}{\lambda} vt = \frac{2\pi}{\lambda}\left(\frac{\lambda}{T}\right)t = \frac{2\pi}{T}t = \omega t.$$

The wave function for a simple harmonic wave on a string reduces to
Equation:

$$y\left(x, t\right) = A\sin\left(kx \mp \omega t\right),$$

where *A* is the amplitude, $k = \frac{2\pi}{\lambda}$ is the wave number, $\omega = \frac{2\pi}{T}$ is the angular frequency, the minus sign is for waves moving in the positive *x*-direction, and the plus sign is for waves moving in the negative *x*-direction. The velocity of the wave is equal to

Note:
Equation:

$$v = \frac{\lambda}{T} = \frac{\lambda}{T}\left(\frac{2\pi}{2\pi}\right) = \frac{\omega}{k}.$$

Think back to our discussion of a mass on a spring, when the position of the mass was modeled as $x\left(t\right) = A\cos\left(\omega t + \phi\right)$. The angle ϕ is a phase shift, added to allow for the fact that the mass may have initial conditions other than $x = +A$ and $v = 0$. For similar reasons, the initial phase is added to the wave function. The wave function modeling a sinusoidal wave, allowing for an initial phase shift ϕ, is

Note:
Equation:

$$y\left(x, t\right) = A\sin\left(kx \mp \omega t + \phi\right)$$

The value

Note:
Equation:

$$(kx \mp \omega t + \phi)$$

is known as the phase of the wave, where ϕ is the initial phase of the wave function. Whether the temporal term ωt is negative or positive depends on the direction of the wave. First consider the minus sign for a wave with an initial phase equal to zero $(\phi = 0)$. The phase of the wave would be $(kx - \omega t)$. Consider following a point on a wave, such as a crest. A crest will occur when $\sin\,(kx - \omega t) = 1.00$, that is, when $kx - \omega t = n\pi + \frac{\pi}{2}$, for any integral value of *n*. For instance, one particular crest occurs at $kx - \omega t = \frac{\pi}{2}$. As the wave moves, time increases and *x* must also increase to keep the phase equal to $\frac{\pi}{2}$. Therefore, the minus sign is for a wave moving in the positive *x*-direction. Using the plus sign, $kx + \omega t = \frac{\pi}{2}$. As time increases, *x* must decrease to keep the phase equal to $\frac{\pi}{2}$. The plus sign is used for waves moving in the negative *x*-direction. In summary, $y\,(x,t) = A\,\sin\,(kx - \omega t + \phi)$ models a wave moving in the positive *x*-direction and $y\,(x,t) = A\,\sin\,(kx + \omega t + \phi)$ models a wave moving in the negative *x*-direction.

[link] is known as a simple harmonic wave function. A wave function is any function such that $f\,(x,t) = f\,(x - vt)$. Later in this chapter, we will see that it is a solution to the linear wave equation. Note that $y\,(x,t) = A\,\cos\,(kx + \omega t + \phi\prime)$ works equally well because it corresponds to a different phase shift $\phi\prime = \phi - \frac{\pi}{2}$.

Note:
Problem-Solving Strategy: Finding the Characteristics of a Sinusoidal Wave

1. To find the amplitude, wavelength, period, and frequency of a sinusoidal wave, write down the wave function in the form $y\,(x,t) = A\,\sin\,(kx - \omega t + \phi)$.
2. The amplitude can be read straight from the equation and is equal to *A*.
3. The period of the wave can be derived from the angular frequency $\left(T = \frac{2\pi}{\omega}\right)$.
4. The frequency can be found using $f = \frac{1}{T}$.
5. The wavelength can be found using the wave number $\left(\lambda = \frac{2\pi}{k}\right)$.

Example:
Characteristics of a Traveling Wave on a String
A transverse wave on a taut string is modeled with the wave function
Equation:

$$y\,(x,t) = A\,\sin\,(kx - wt) = 0.2\,\text{m}\,\sin\left(6.28\,\text{m}^{-1}x - 1.57\,\text{s}^{-1}t\right).$$

Find the amplitude, wavelength, period, and speed of the wave.
Strategy
All these characteristics of the wave can be found from the constants included in the equation or from simple combinations of these constants.
Solution

1. The amplitude, wave number, and angular frequency can be read directly from the wave equation:
   Equation:

$$y\left(x,t\right)=A\sin\left(kx-wt\right)=0.2\,\text{m}\sin\left(6.28\,\text{m}^{-1}x-1.57\,\text{s}^{-1}t\right).$$

Equation:

$$\left(A=0.2\,\text{m};\,k=6.28\,\text{m}^{-1};\,\omega=1.57\,\text{s}^{-1}\right)$$

2. The wave number can be used to find the wavelength:
   Equation:

$$\begin{array}{c}k=\frac{2\pi}{\lambda}.\\ \lambda=\frac{2\pi}{k}=\frac{2\pi}{6.28\,\text{m}^{-1}}=1.0\,\text{m}.\end{array}$$

3. The period of the wave can be found using the angular frequency:
   Equation:

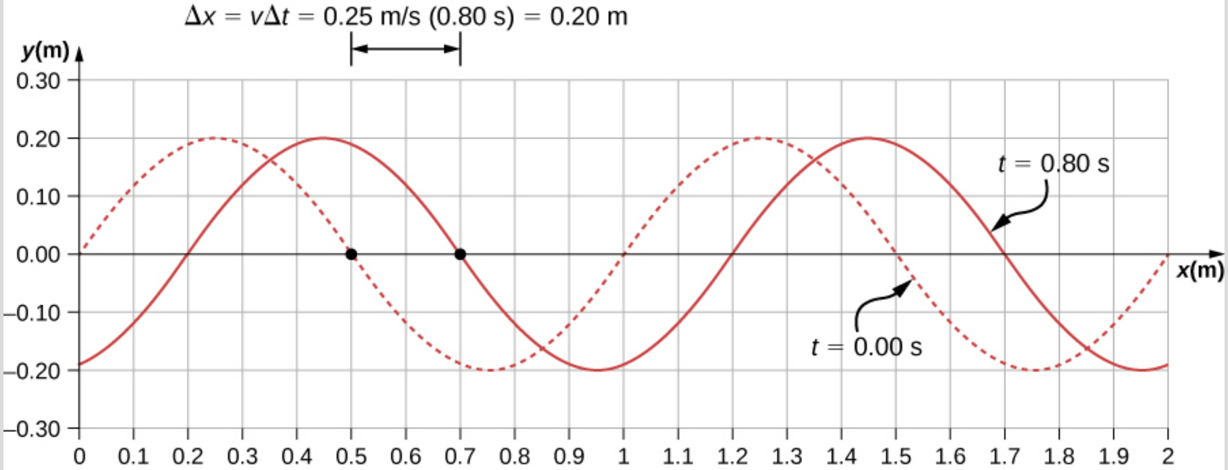
$$\begin{array}{c}\omega=\frac{2\pi}{T}.\\ T=\frac{2\pi}{\omega}=\frac{2\pi}{1.57\,\text{s}^{-1}}=4\,\text{s}.\end{array}$$

4. The speed of the wave can be found using the wave number and the angular frequency. The direction of the wave can be determined by considering the sign of $kx\mp\omega t$: A negative sign suggests that the wave is moving in the positive *x*-direction:
   Equation:

$$|v|=\frac{\omega}{k}=\frac{1.57\,\text{s}^{-1}}{6.28\,\text{m}^{-1}}=0.25\,\text{m/s}.$$

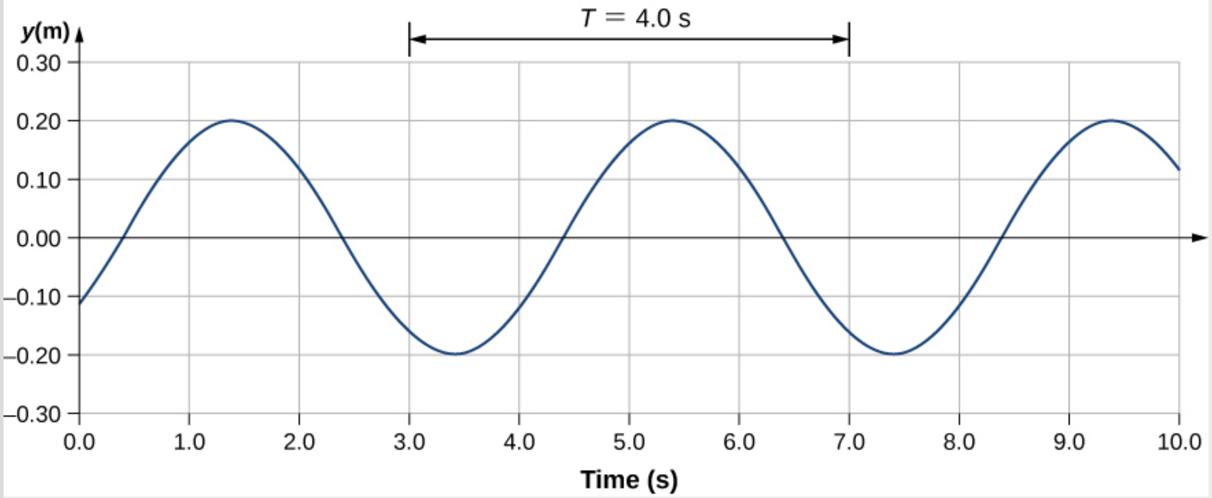
Significance

All of the characteristics of the wave are contained in the wave function. Note that the wave speed is the speed of the wave in the direction parallel to the motion of the wave. Plotting the height of the medium *y* versus the position *x* for two times $t=0.00\,\text{s}$ and $t=0.80\,\text{s}$ can provide a graphical visualization of the wave ([link]).



A graph of height of the wave *y* as a function of position *x* for snapshots of the wave at two times. The dotted line represents the wave at time $t=0.00\,\text{s}$ and the solid line represents the wave at $t=0.80\,\text{s}$. Since the wave velocity is constant, the distance the wave travels is the wave velocity times the time interval. The black dots indicate the points used to measure the displacement of the wave. The medium moves up and down, whereas the wave moves to the right.

There is a second velocity to the motion. In this example, the wave is transverse, moving horizontally as the medium oscillates up and down perpendicular to the direction of motion. The graph in [link] shows the motion of the medium at point $x = 0.60 \text{ m}$ as a function of time. Notice that the medium of the wave oscillates up and down between $y = +0.20 \text{ m}$ and $y = -0.20 \text{ m}$ every period of 4.0 seconds.



A graph of height of the wave *y* as a function of time *t* for the position $x = 0.6 \text{ m}$. The medium oscillates between $y = +0.20 \text{ m}$ and $y = -0.20 \text{ m}$ every period. The period represented picks two convenient points in the oscillations to measure the period. The period can be measured between any two adjacent points with the same amplitude and the same velocity, $(\partial y / \partial t)$. The velocity can be found by looking at the slope tangent to the point on a *y*-versus-*t* plot. Notice that at times $t = 3.00 \; s$ and $t = 7.00 s$, the heights and the velocities are the same and the period of the oscillation is 4.00 s.

Note:
Exercise:

Problem:

Check Your Understanding The wave function above is derived using a sine function. Can a cosine function be used instead?

Solution:

Yes, a cosine function is equal to a sine function with a phase shift, and either function can be used in a wave function. Which function is more convenient to use depends on the initial conditions. In [link], the wave has an initial height of $\text{y}(0.00, 0.00) = 0$ and then the wave height increases to the maximum height at the crest. If the initial height at the initial time was equal to the amplitude of the wave $\text{y}(0.00, 0.00) = +A$, then it might be more convenient to model the wave with a cosine function.

Velocity and Acceleration of the Medium

As seen in [link], the wave speed is constant and represents the speed of the wave as it propagates through the medium, not the speed of the particles that make up the medium. The particles of the medium oscillate around an equilibrium position as the wave propagates through the medium. In the case of the transverse wave propagating in the *x*-direction, the particles oscillate up and down in the *y*-direction, perpendicular to the motion of the wave. The velocity of the particles of the medium is not constant, which means there is an acceleration. The velocity of the medium, which is perpendicular to the wave velocity in a transverse wave, can be found by taking the partial derivative of the position equation with respect to time. The partial derivative is found by taking the derivative of the function, treating all variables as constants, except for the variable in question. In the case of the partial derivative with respect to time *t*, the position *x* is treated as a constant. Although this may sound strange if you haven't seen it before, the object of this exercise is to find the transverse velocity at a point, so in this sense, the *x*-position is not changing. We have

Equation:

$$\begin{aligned} y(x,t) &= A\sin(kx-\omega t+\phi) \\ v_y(x,t) &= \tfrac{\partial y(x,t)}{\partial t} = \tfrac{\partial}{\partial t}(A\sin(kx-\omega t+\phi)) \\ &= -A\omega\cos(kx-\omega t+\phi) \\ &= -v_{y\,\max}\cos(kx-\omega t+\phi). \end{aligned}$$

The magnitude of the maximum velocity of the medium is $|v_{y_{\max}}| = A\omega$. This may look familiar from the Oscillations and a mass on a spring.

We can find the acceleration of the medium by taking the partial derivative of the velocity equation with respect to time,

Equation:

$$\begin{aligned} a_y(x,t) &= \tfrac{\partial v_y}{\partial t} = \tfrac{\partial}{\partial t}(-A\omega\cos(kx-\omega t+\phi)) \\ &= -A\omega^2\sin(kx-\omega t+\phi) \\ &= -a_{y\,\max}\sin(kx-\omega t+\phi). \end{aligned}$$

The magnitude of the maximum acceleration is $|a_{y_{\max}}| = A\omega^2$. The particles of the medium, or the mass elements, oscillate in simple harmonic motion for a mechanical wave.

The Linear Wave Equation

We have just determined the velocity of the medium at a position *x* by taking the partial derivative, with respect to time, of the position *y*. For a transverse wave, this velocity is perpendicular to the direction of propagation of the wave. We found the acceleration by taking the partial derivative, with respect to time, of the velocity, which is the second time derivative of the position:

Equation:

$$a_y(x,t) = \frac{\partial^2 y(x,t)}{\partial t^2} = \frac{\partial^2}{\partial t^2}(A\sin(kx-\omega t+\phi)) = -A\omega^2\sin(kx-\omega t+\phi).$$

Now consider the partial derivatives with respect to the other variable, the position *x*, holding the time constant. The first derivative is the slope of the wave at a point *x* at a time *t*,

Equation:

$$\text{slope} = \frac{\partial y(x,t)}{\partial x} = \frac{\partial}{\partial x}(A\sin(kx-\omega t+\phi)) = Ak\cos(kx-\omega t+\phi).$$

The second partial derivative expresses how the slope of the wave changes with respect to position—in other words, the curvature of the wave, where

Equation:

$$\text{curvature} = \frac{\partial^2 y(x,t)}{\partial x^2} = \frac{\partial^2}{\partial^2 x}\left(A\sin(kx-\omega t+\phi)\right) = -Ak^2\sin(kx-\omega t+\phi).$$

The ratio of the acceleration and the curvature leads to a very important relationship in physics known as the **linear wave equation**. Taking the ratio and using the equation $v = \omega/k$ yields the linear wave equation (also known simply as the wave equation or the equation of a vibrating string),

Equation:

$$\begin{aligned} \frac{\frac{\partial^2 y(x,t)}{\partial t^2}}{\frac{\partial^2 y(x,t)}{\partial x^2}} &= \frac{-A\omega^2\sin(kx-\omega t+\phi)}{-Ak^2\sin(kx-\omega t+\phi)} \\ &= \frac{\omega^2}{k^2} = v^2, \end{aligned}$$

Note:

Equation:

$$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y(x,t)}{\partial t^2}.$$

[link] is the linear wave equation, which is one of the most important equations in physics and engineering. We derived it here for a transverse wave, but it is equally important when investigating longitudinal waves. This relationship was also derived using a sinusoidal wave, but it successfully describes any wave or pulse that has the form $y(x,t) = f(x \mp vt)$. These waves result due to a linear restoring force of the medium—thus, the name linear wave equation. Any wave function that satisfies this equation is a linear wave function.

An interesting aspect of the linear wave equation is that if two wave functions are individually solutions to the linear wave equation, then the sum of the two linear wave functions is also a solution to the wave equation. Consider two transverse waves that propagate along the *x*-axis, occupying the same medium. Assume that the individual waves can be modeled with the wave functions $y_1(x,t) = f(x \mp vt)$ and $y_2(x,t) = g(x \mp vt)$, which are solutions to the linear wave equations and are therefore linear wave functions. The sum of the wave functions is the wave function

Equation:

$$y_1(x,t) + y_2(x,t) = f(x \mp vt) + g(x \mp vt).$$

Consider the linear wave equation:

Equation:

$$\begin{aligned} \frac{\partial^2(f+g)}{\partial x^2} &= \frac{1}{v^2}\frac{\partial^2(f+g)}{\partial t^2} \\ \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 g}{\partial x^2} &= \frac{1}{v^2}\left[\frac{\partial^2 f}{\partial t^2} + \frac{\partial^2 g}{\partial t^2}\right]. \end{aligned}$$

This has shown that if two linear wave functions are added algebraically, the resulting wave function is also linear. This wave function models the displacement of the medium of the resulting wave at each position along the *x*-axis. If two linear waves occupy the same medium, they are said to interfere. If these waves can be modeled with a linear wave function, these wave functions add to form the wave equation of the wave resulting from the interference of the individual waves. The displacement of the medium at every point of the resulting wave is the algebraic sum of the displacements due to the individual waves.

Taking this analysis a step further, if wave functions $y_1(x,t) = f(x \mp vt)$ and $y_2(x,t) = g(x \mp vt)$ are solutions to the linear wave equation, then $Ay_1(x,t) + By_2(x,y)$, where *A* and *B* are constants, is also a solution to the linear wave equation. This property is known as the principle of superposition. Interference and superposition are covered in more detail in Interference of Waves.

Example:

Interference of Waves on a String

Consider a very long string held taut by two students, one on each end. Student A oscillates the end of the string producing a wave modeled with the wave function $y_1(x,t) = A\sin(kx - \omega t)$ and student B oscillates the string producing at twice the frequency, moving in the opposite direction. Both waves move at the same speed $v = \frac{\omega}{k}$. The two waves interfere to form a resulting wave whose wave function is $y_R(x,t) = y_1(x,t) + y_2(x,t)$. Find the velocity of the resulting wave using the linear wave equation $\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y(x,t)}{\partial t^2}$.

Strategy

First, write the wave function for the wave created by the second student. Note that the angular frequency of the second wave is twice the frequency of the first wave (2ω), and since the velocity of the two waves are the same, the wave number of the second wave is twice that of the first wave $(2k)$. Next, write the wave equation for the resulting wave function, which is the sum of the two individual wave functions. Then find the second partial derivative with respect to position and the second partial derivative with respect to time. Use the linear wave equation to find the velocity of the resulting wave.

Solution

1. Write the wave function of the second wave: $y_2(x,t) = A\sin(2kx + 2\omega t)$.
2. Write the resulting wave function:
   Equation:

$$y_R(x,t) = y_1(x,t) + y(x,t) = A\sin(kx - \omega t) + A\sin(2kx + 2\omega t).$$

3. Find the partial derivatives:
   Equation:

$$\begin{aligned}
\frac{\partial y_R(x,t)}{\partial x} &= -Ak\cos(kx - \omega t) + 2Ak\cos(2kx + 2\omega t),\\
\frac{\partial^2 y_R(x,t)}{\partial^2 x} &= -Ak^2\sin(kx - \omega t) - 4Ak^2\sin(2kx + 2\omega t),\\
\frac{\partial y_R(x,t)}{\partial t} &= -A\omega\cos(kx - \omega t) + 2A\omega\cos(2kx + 2\omega t),\\
\frac{\partial^2 y_R(x,t)}{\partial^2 t} &= -A\omega^2\sin(kx - \omega t) - 4A\omega^2\sin(2kx + 2\omega t).
\end{aligned}$$

4. Use the wave equation to find the velocity of the resulting wave:

$$\begin{aligned} \frac{\partial^2 y(x,t)}{\partial x^2} &= \frac{1}{v^2}\,\frac{\partial^2 y(x,t)}{\partial t^2}, \\ -Ak^2 \sin\left(kx-\omega t\right) - 4Ak^2 \sin\left(2kx+2\omega t\right) &= \frac{1}{v^2}\left(-A\omega^2 \sin\left(kx-\omega t\right) - 4A\omega^2 \sin\left(2kx+2\omega t\right)\right), \\ k^2\left(-A\sin\left(kx-\omega t\right) - 4A\sin\left(2kx+2\omega t\right)\right) &= \frac{\omega^2}{v^2}\left(-A\sin\left(kx-\omega t\right) - 4A\sin\left(2kx+2\omega t\right)\right), \\ k^2 &= \frac{\omega^2}{v^2},\ |v| = \frac{\omega}{k}. \end{aligned}$$

Significance
The speed of the resulting wave is equal to the speed of the original waves $\left(v = \frac{\omega}{k}\right)$. We will show in the next section that the speed of a simple harmonic wave on a string depends on the tension in the string and the mass per length of the string. For this reason, it is not surprising that the component waves as well as the resultant wave all travel at the same speed.

Note:
Exercise:

Problem:

Check Your Understanding The wave equation $\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2}\,\frac{\partial^2 y(x,t)}{\partial t^2}$ works for any wave of the form $y\left(x,t\right) = f\left(x \mp vt\right)$. In the previous section, we stated that a cosine function could also be used to model a simple harmonic mechanical wave. Check if the wave
Equation:

$$y\left(x,t\right) = 0.50\,\mathrm{m}\cos\left(0.20\pi\,\mathrm{m}^{-1}x - 4.00\pi\,\mathrm{s}^{-1}t + \frac{\pi}{10}\right)$$

is a solution to the wave equation.

Solution:

This wave, with amplitude $A = 0.5\,\mathrm{m}$, wavelength $\lambda = 10.00\,\mathrm{m}$, period $T = 0.50\,\mathrm{s}$, is a solution to the wave equation with a wave velocity $v = 20.00\,\mathrm{m/s}$.

Any disturbance that complies with the wave equation can propagate as a wave moving along the *x*-axis with a wave speed *v*. It works equally well for waves on a string, sound waves, and electromagnetic waves. This equation is extremely useful. For example, it can be used to show that electromagnetic waves move at the speed of light.

Summary

- A wave is an oscillation (of a physical quantity) that travels through a medium, accompanied by a transfer of energy. Energy transfers from one point to another in the direction of the wave motion. The particles of the medium oscillate up and down, back and forth, or both up and down and back and forth, around an equilibrium position.
- A snapshot of a sinusoidal wave at time $t = 0.00\,\mathrm{s}$ can be modeled as a function of position. Two examples of such functions are $y\left(x\right) = A\sin\left(kx+\phi\right)$ and $y\left(x\right) = A\cos\left(kx+\phi\right)$.
- Given a function of a wave that is a snapshot of the wave, and is only a function of the position *x*, the motion of the pulse or wave moving at a constant velocity can be modeled with the function, replacing *x* with $x \mp vt$. The minus sign is for motion in the positive direction and the plus sign for the negative direction.

- The wave function is given by $y\left(x,t\right)=A\sin\left(kx-\omega t+\phi\right)$ where $k=2\pi/\lambda$ is defined as the wave number, $\omega=2\pi/T$ is the angular frequency, and ϕ is the phase shift.
- The wave moves with a constant velocity v_w, where the particles of the medium oscillate about an equilibrium position. The constant velocity of a wave can be found by $v=\frac{\lambda}{T}=\frac{\omega}{k}$.

Conceptual Questions

Exercise:

Problem:

If you were to shake the end of a taut spring up and down 10 times a second, what would be the frequency and the period of the sinusoidal wave produced on the spring?

Exercise:

Problem:

If you shake the end of a stretched spring up and down with a frequency *f*, you can produce a sinusoidal, transverse wave propagating down the spring. Does the wave number depend on the frequency you are shaking the spring?

Solution:

The wavelength is equal to the velocity of the wave times the frequency and the wave number is equal to $k=\frac{2\pi}{\lambda}$, so yes, the wave number will depend on the frequency and also depend on the velocity of the wave propagating through the spring.

Exercise:

Problem:

Does the vertical speed of a segment of a horizontal taut string through which a sinusoidal, transverse wave is propagating depend on the wave speed of the transverse wave?

Exercise:

Problem:

In this section, we have considered waves that move at a constant wave speed. Does the medium accelerate?

Solution:

The medium moves in simple harmonic motion as the wave propagates through the medium, continuously changing speed, therefore it accelerates. The acceleration of the medium is due to the restoring force of the medium, which acts in the opposite direction of the displacement.

Exercise:

Problem:

If you drop a pebble in a pond you may notice that several concentric ripples are produced, not just a single ripple. Why do you think that is?

Problems

Exercise:

Problem:

A pulse can be described as a single wave disturbance that moves through a medium. Consider a pulse that is defined at time $t = 0.00\text{ s}$ by the equation $y\left(x\right) = \frac{6.00\text{ m}^3}{x^2 + 2.00\text{ m}^2}$ centered around $x = 0.00\text{ m}$. The pulse moves with a velocity of $v = 3.00\text{ m/s}$ in the positive *x*-direction. (a) What is the amplitude of the pulse? (b) What is the equation of the pulse as a function of position and time? (c) Where is the pulse centered at time $t = 5.00\text{ s}$?

Exercise:

Problem:

A transverse wave on a string is modeled with the wave function $y\left(x,t\right) = \left(0.20\text{ cm}\right)\text{sin}\left(2.00\text{ m}^{-1}x - 3.00\text{ s}^{-1}t + \frac{\pi}{16}\right)$. What is the height of the string with respect to the equilibrium position at a position $x = 4.00\text{ m}$ and a time $t = 10.00\text{ s}$?

Solution:

$y\left(x,t\right) = -0.037\text{ cm}$

Exercise:

Problem:

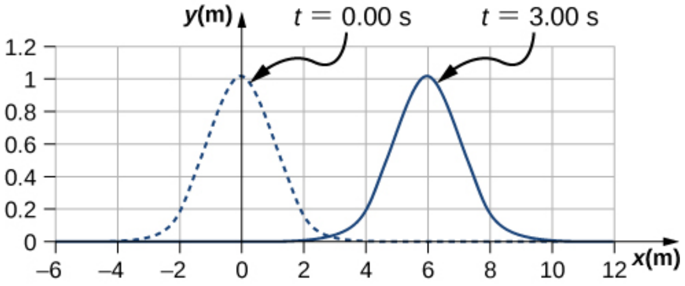
Consider the wave function $y\left(x,t\right) = \left(3.00\text{ cm}\right)\text{sin}\left(0.4\text{ m}^{-1}x + 2.00\text{ s}^{-1}t + \frac{\pi}{10}\right)$. What are the period, wavelength, speed, and initial phase shift of the wave modeled by the wave function?

Exercise:

Problem:

A pulse is defined as $y\left(x,t\right) = e^{-2.77\left(\frac{2.00\left(x - 2.00\text{ m/s}\left(t\right)\right)}{5.00\text{ m}}\right)^2}$. Use a spreadsheet, or other computer program, to plot the pulse as the height of medium *y* as a function of position *x*. Plot the pulse at times $t = 0.00\text{ s}$ and $t = 3.00\text{ s}$ on the same graph. Where is the pulse centered at time $t = 3.00\text{ s}$? Use your spreadsheet to check your answer.

Solution:



The pulse will move $\Delta x = 6.00\text{ m}$.

Exercise:

Problem:

A wave is modeled at time $t = 0.00\text{ s}$ with a wave function that depends on position. The equation is $y\left(x\right) = \left(0.30\text{ m}\right)\text{sin}\left(6.28\text{ m}^{-1}x\right)$. The wave travels a distance of 4.00 meters in 0.50 s in the positive *x*-direction. Write an equation for the wave as a function of position and time.

Exercise:

Problem:

A wave is modeled with the function $y\left(x,t\right)=\left(0.25\,\text{m}\right)\text{cos}\left(0.30\,\text{m}^{-1}x-0.90\,\text{s}^{-1}t+\frac{\pi}{3}\right)$. Find the (a) amplitude, (b) wave number, (c) angular frequency, (d) wave speed, (e) phase shift, (f) wavelength, and (g) period of the wave.

Solution:

a. $A=0.25\,\text{m}$; b. $k=0.30\,\text{m}^{-1}$; c. $\omega=0.90\,\text{s}^{-1}$; d. $v=3.0\,\text{m/s}$; e. $\phi=\pi/3\,\text{rad}$; f. $\lambda=20.93\,\text{m}$; g. $T=6.98\,\text{s}$

Exercise:

Problem:

A surface ocean wave has an amplitude of 0.60 m and the distance from trough to trough is 8.00 m. It moves at a constant wave speed of 1.50 m/s propagating in the positive *x*-direction. At $t=0$, the water displacement at $x=0$ is zero, and v_y is positive. (a) Assuming the wave can be modeled as a sine wave, write a wave function to model the wave. (b) Use a spreadsheet to plot the wave function at times $t=0.00\,\text{s}$ and $t=2.00\,\text{s}$ on the same graph. Verify that the wave moves 3.00 m in those 2.00 s.

Exercise:

Problem:

A wave is modeled by the wave function $y\left(x,t\right)=\left(0.30\,\text{m}\right)\text{sin}\left[\frac{2\pi}{4.50\,\text{m}}\left(x-18.00\,\frac{\text{m}}{\text{s}}t\right)\right]$. What are the amplitude, wavelength, wave speed, period, and frequency of the wave?

Solution:

$A=0.30\,\text{m},\ \lambda=4.50\,\text{m},\ v=18.00\,\text{m/s},\ f=4.00\,\text{Hz},\ T=0.25\,\text{s}$

Exercise:

Problem:

A transverse wave on a string is described with the wave function $y\left(x,t\right)=\left(0.50\,\text{cm}\right)\text{sin}\left(1.57\,\text{m}^{-1}x-6.28\,\text{s}^{-1}t\right)$. (a) What is the wave velocity of the wave? (b) What is the magnitude of the maximum velocity of the string perpendicular to the direction of the motion?

Exercise:

Problem:

A swimmer in the ocean observes one day that the ocean surface waves are periodic and resemble a sine wave. The swimmer estimates that the vertical distance between the crest and the trough of each wave is approximately 0.45 m, and the distance between each crest is approximately 1.8 m. The swimmer counts that 12 waves pass every two minutes. Determine the simple harmonic wave function that would describes these waves.

Solution:

$y\left(x,t\right)=0.23\,\text{m}\,\text{sin}\left(3.49\,\text{m}^{-1}x-0.63\,\text{s}^{-1}t\right)$

Exercise:

Problem:

Consider a wave described by the wave function $y\left(x,t\right)=0.3\,\text{m}\,\text{sin}\left(2.00\,{\text{m}}^{-1}x-628.00\,{\text{s}}^{-1}t\right).$ (a) How many crests pass by an observer at a fixed location in 2.00 minutes? (b) How far has the wave traveled in that time?

Exercise:

Problem:

Consider two waves defined by the wave functions ${y}_{1}\left(x,t\right)=0.50\,\text{m}\,\text{sin}\left(\frac{2\pi}{3.00\,\text{m}}x+\frac{2\pi}{4.00\,\text{s}}t\right)$ and ${y}_{2}\left(x,t\right)=0.50\,\text{m}\,\text{sin}\left(\frac{2\pi}{6.00\,\text{m}}x-\frac{2\pi}{4.00\,\text{s}}t\right).$ What are the similarities and differences between the two waves?

Solution:

They have the same angular frequency, frequency, and period. They are traveling in opposite directions and ${y}_{2}\left(x,t\right)$ has twice the wavelength as ${y}_{1}\left(x,t\right)$ and is moving at half the wave speed.

Exercise:

Problem:

Consider two waves defined by the wave functions ${y}_{1}\left(x,t\right)=0.20\,\text{m}\,\text{sin}\left(\frac{2\pi}{6.00\,\text{m}}x-\frac{2\pi}{4.00\,\text{s}}t\right)$ and ${y}_{2}\left(x,t\right)=0.20\,\text{m}\,\text{cos}\left(\frac{2\pi}{6.00\,\text{m}}x-\frac{2\pi}{4.00\,\text{s}}t\right).$ What are the similarities and differences between the two waves?

Exercise:

Problem:

The speed of a transverse wave on a string is 300.00 m/s, its wavelength is 0.50 m, and the amplitude is 20.00 cm. How much time is required for a particle on the string to move through a distance of 5.00 km?

Solution:

Each particle of the medium moves a distance of 4*A* each period. The period can be found by dividing the velocity by the wavelength: $t=10.42\,\text{s}$

Glossary

linear wave equation	equation describing waves that result from a linear restoring force of the medium; any function that is a solution to the wave equation describes a wave moving in the positive *x*-direction or the negative *x*-direction with a constant wave speed *v*
pulse	single disturbance that moves through a medium, transferring energy but not mass
wave function	mathematical model of the position of particles of the medium
wave number	$\frac{2\pi}{\lambda}$

Wave Speed on a Stretched String

By the end of this section, you will be able to:

- Determine the factors that affect the speed of a wave on a string
- Write a mathematical expression for the speed of a wave on a string and generalize these concepts for other media

The speed of a wave depends on the characteristics of the medium. For example, in the case of a guitar, the strings vibrate to produce the sound. The speed of the waves on the strings, and the wavelength, determine the frequency of the sound produced. The strings on a guitar have different thickness but may be made of similar material. They have different *linear densities*, where the linear density is defined as the mass per length,

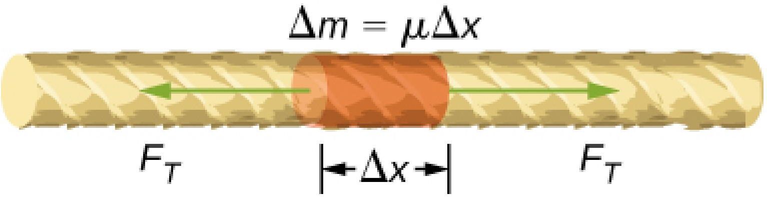
Note:
Equation:

$$\mu = \frac{\text{mass of string}}{\text{length of string}} = \frac{m}{l}.$$

In this chapter, we consider only string with a constant linear density. If the linear density is constant, then the mass (Δm) of a small length of string (Δx) is $\Delta m = \mu \Delta x$. For example, if the string has a length of 2.00 m and a mass of 0.06 kg, then the linear density is $\mu = \frac{0.06\,\text{kg}}{2.00\,\text{m}} = 0.03\frac{\text{kg}}{\text{m}}$. If a 1.00-mm section is cut from the string, the mass of the 1.00-mm length is $\Delta m = \mu \Delta x = \left(0.03\frac{\text{kg}}{\text{m}}\right)0.001\,\text{m} = 3.00 \times 10^{-5}\,\text{kg}$. The guitar also has a method to change the tension of the strings. The tension of the strings is adjusted by turning spindles, called the tuning pegs, around which the strings are wrapped. For the guitar, the linear density of the string and the tension in the string determine the speed of the waves in the string and the frequency of the sound produced is proportional to the wave speed.

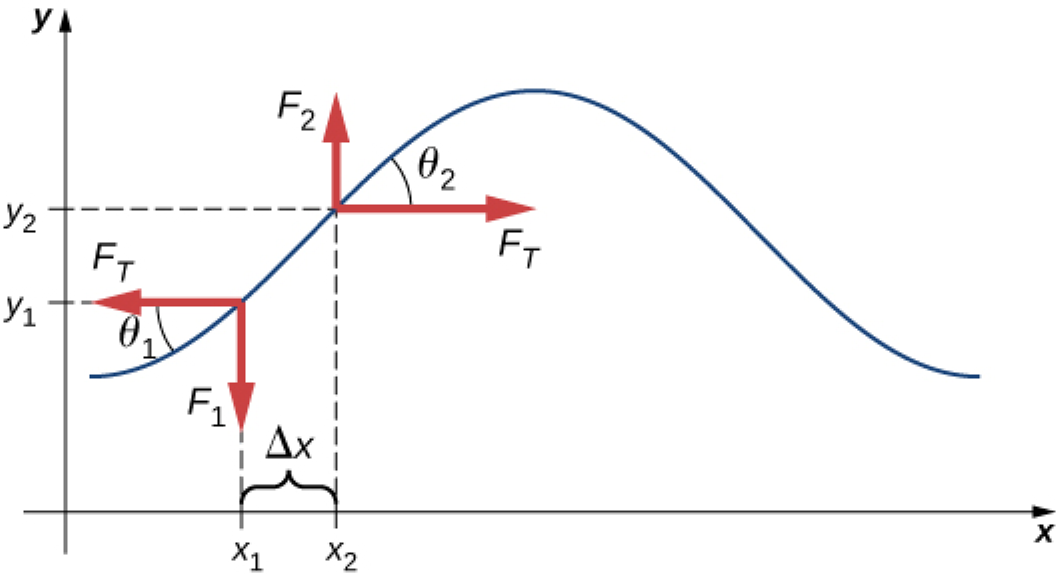
Wave Speed on a String under Tension

To see how the speed of a wave on a string depends on the tension and the linear density, consider a pulse sent down a taut string ([link]). When the taut string is at rest at the equilibrium position, the tension in the string F_T is constant. Consider a small element of the string with a mass equal to $\Delta m = \mu \Delta x$. The mass element is at rest and in equilibrium and the force of tension of either side of the mass element is equal and opposite.



Mass element of a string kept taut with a tension F_T. The mass element is in static equilibrium, and the force of tension acting on either side of the mass element is equal in magnitude and opposite in direction.

If you pluck a string under tension, a transverse wave moves in the positive *x*-direction, as shown in [link]. The mass element is small but is enlarged in the figure to make it visible. The small mass element oscillates perpendicular to the wave motion as a result of the restoring force provided by the string and does not move in the *x*-direction. The tension F_T in the string, which acts in the positive and negative *x*-direction, is approximately constant and is independent of position and time.



A string under tension is plucked, causing a pulse to move along the string in the positive *x*-direction.

Assume that the inclination of the displaced string with respect to the horizontal axis is small. The net force on the element of the string, acting parallel to the string, is the sum of the tension in the string and the restoring force. The *x*-components of the force of tension cancel, so the net force is equal to the sum of the *y*-components of the force. The magnitude of the *x*-component of the force is equal to the horizontal force of tension of the string F_T as shown in [link]. To obtain the *y*-components of the force, note that $\tan\theta_1 = \frac{-F_1}{F_T}$ and $\tan\theta_2 = \frac{F_2}{F_T}$. The $\tan\theta$ is equal to the slope of a function at a point, which is equal to the partial derivative of *y* with respect to *x* at that point. Therefore, $\frac{F_1}{F_T}$ is equal to the negative slope of the string at x_1 and $\frac{F_2}{F_T}$ is equal to the slope of the string at x_2 :

Equation:

$$\frac{F_1}{F_T} = -\left(\frac{\partial y}{\partial x}\right)_{x_1} \text{ and } \frac{F_2}{F_T} = -\left(\frac{\partial y}{\partial x}\right)_{x_2}.$$

The net force is on the small mass element can be written as

Equation:

$$F_{\text{net}} = F_1 + F_2 = F_T\left[\left(\frac{\partial y}{\partial x}\right)_{x_2} - \left(\frac{\partial y}{\partial x}\right)_{x_1}\right].$$

Using Newton’s second law, the net force is equal to the mass times the acceleration. The linear density of the string μ is the mass per length of the string, and the mass of the portion of the string is $\mu\Delta x$,

Equation:

$$\begin{aligned} F_T\left[\left(\tfrac{\partial y}{\partial x}\right)_{x_2} - \left(\tfrac{\partial y}{\partial x}\right)_{x_1}\right] &= \Delta ma, \\ F_T\left[\left(\tfrac{\partial y}{\partial x}\right)_{x_2} - \left(\tfrac{\partial y}{\partial x}\right)_{x_1}\right] &= \mu\Delta x\,\frac{\partial^2 y}{\partial t^2}. \end{aligned}$$

Dividing by $F_T\Delta x$ and taking the limit as Δx approaches zero,

Equation:

$$\begin{aligned} \frac{\left[\left(\frac{\partial y}{\partial x}\right)_{x_2} - \left(\frac{\partial y}{\partial x}\right)_{x_1}\right]}{\Delta x} &= \frac{\mu}{F_T}\,\frac{\partial^2 y}{\partial t^2} \\ \lim_{\Delta x\to 0}\frac{\left[\left(\frac{\partial y}{\partial x}\right)_{x_2} - \left(\frac{\partial y}{\partial x}\right)_{x_1}\right]}{\Delta x} &= \frac{\mu}{F_T}\,\frac{\partial^2 y}{\partial t^2} \\ \frac{\partial^2 y}{\partial x^2} = \frac{\mu}{F_T}\,\frac{\partial^2 y}{\partial t^2}. \end{aligned}$$

Recall that the linear wave equation is

Equation:

$$\frac{\partial^2 y\left(x,t\right)}{\partial x^2} = \frac{1}{v^2}\,\frac{\partial^2 y\left(x,t\right)}{\partial t^2}.$$

Therefore,

Equation:

$$\frac{1}{v^2} = \frac{\mu}{F_T}.$$

Solving for *v*, we see that the speed of the wave on a string depends on the tension and the linear density.

Note:

Speed of a Wave on a String Under Tension

The speed of a pulse or wave on a string under tension can be found with the equation

Equation:

$$|v| = \sqrt{\frac{F_T}{\mu}}$$

where F_T is the tension in the string and μ is the mass per length of the string.

Example:

The Wave Speed of a Guitar Spring

On a six-string guitar, the high E string has a linear density of $\mu_{\text{High E}} = 3.09\ \times\ 10^{-4}\,\text{kg/m}$ and the low E string has a linear density of $\mu_{\text{Low E}} = 5.78\ \times\ 10^{-3}\,\text{kg/m}.$ (a) If the high E string is plucked, producing a wave in the string, what is the speed of the wave if the tension of the string is 56.40 N? (b) The linear density of the low E string is approximately 20 times greater than that of the high E string. For waves to travel through the low E string at the same wave speed as the high E, would the tension need to be larger or smaller than the high E string? What would be the approximate tension? (c) Calculate the tension of the low E string needed for the same wave speed.

Strategy

a. The speed of the wave can be found from the linear density and the tension $v = \sqrt{\frac{F_T}{\mu}}$.
b. From the equation $v = \sqrt{\frac{F_T}{\mu}}$, if the linear density is increased by a factor of almost 20, the tension would need to be increased by a factor of 20.
c. Knowing the velocity and the linear density, the velocity equation can be solved for the force of tension $F_T = \mu v^2$.

Solution

a. Use the velocity equation to find the speed:
   Equation:

$$v = \sqrt{\frac{F_T}{\mu}} = \sqrt{\frac{56.40\,\text{N}}{3.09 \times 10^{-4}\,\text{kg/m}}} = 427.23\,\text{m/s}.$$

b. The tension would need to be increased by a factor of approximately 20. The tension would be slightly less than 1128 N.
c. Use the velocity equation to find the actual tension:
   Equation:

$$F_T = \mu v^2 = 5.78 \times 10^{-3}\text{kg/m}(427.23\,\text{m/s})^2 = 1055.00\,\text{N}.$$

This solution is within 7% of the approximation.

Significance

The standard notes of the six string (high E, B, G, D, A, low E) are tuned to vibrate at the fundamental frequencies (329.63 Hz, 246.94Hz, 196.00Hz, 146.83Hz, 110.00Hz, and 82.41Hz) when plucked. The frequencies depend on the speed of the waves on the string and the wavelength of the waves. The six strings have different linear densities and are “tuned” by changing the tensions in the strings. We will see in Interference of Waves that the wavelength depends on the length of the strings and the boundary conditions. To play notes other than the fundamental notes, the lengths of the strings are changed by pressing down on the strings.

Note:
Exercise:

Problem:

Check Your Understanding The wave speed of a wave on a string depends on the tension and the linear mass density. If the tension is doubled, what happens to the speed of the waves on the string?

Solution:

Since the speed of a wave on a taunt string is proportional to the square root of the tension divided by the linear density, the wave speed would increase by $\sqrt{2}$.

Speed of Compression Waves in a Fluid

The speed of a wave on a string depends on the square root of the tension divided by the mass per length, the linear density. In general, the speed of a wave through a medium depends on the elastic property of the medium and the inertial property of the medium.

Equation:

$$|v| = \sqrt{\frac{\text{elastic property}}{\text{inertial property}}}$$

The elastic property describes the tendency of the particles of the medium to return to their initial position when perturbed. The inertial property describes the tendency of the particle to resist changes in velocity.

The speed of a longitudinal wave through a liquid or gas depends on the density of the fluid and the bulk modulus of the fluid,

Note:
Equation:

$$v=\sqrt{\frac{B}{\rho}}.$$

Here the bulk modulus is defined as $B=-\frac{\Delta P}{\Delta V/V_0}$, where ΔP is the change in the pressure and the denominator is the ratio of the change in volume to the initial volume, and $\rho\equiv\frac{m}{V}$ is the mass per unit volume. For example, sound is a mechanical wave that travels through a fluid or a solid. The speed of sound in air with an atmospheric pressure of $1.013\ \times\ 10^5\,\mathrm{Pa}$ and a temperature of $20^\circ\mathrm{C}$ is $v_\mathrm{s}\approx343.00\,\mathrm{m/s}$. Because the density depends on temperature, the speed of sound in air depends on the temperature of the air. This will be discussed in detail in Sound.

Summary

- The speed of a wave on a string depends on the linear density of the string and the tension in the string. The linear density is mass per unit length of the string.
- In general, the speed of a wave depends on the square root of the ratio of the elastic property to the inertial property of the medium.
- The speed of a wave through a fluid is equal to the square root of the ratio of the bulk modulus of the fluid to the density of the fluid.

- The speed of sound through air at $T = 20^\circ\text{C}$ is approximately $v_\text{s} = 343.00\,\text{m/s}$.

Conceptual Questions

Exercise:

Problem:

If the tension in a string were increased by a factor of four, by what factor would the wave speed of a wave on the string increase?

Solution:

The wave speed is proportional to the square root of the tension, so the speed is doubled.

Exercise:

Problem:

Does a sound wave move faster in seawater or fresh water, if both the sea water and fresh water are at the same temperature and the sound wave moves near the surface? $\left(\rho_\text{w} \approx 1000\frac{\text{kg}}{\text{m}^3}, \rho_\text{s} \approx 1030\frac{\text{kg}}{\text{m}^3}, B_\text{w} = 2.15 \times 10^9\,\text{Pa}, B_\text{s} = 2.34 \times 10^9\,\text{Pa}\right)$

Exercise:

Problem:

Guitars have strings of different linear mass density. If the lowest density string and the highest density string are under the same tension, which string would support waves with the higher wave speed?

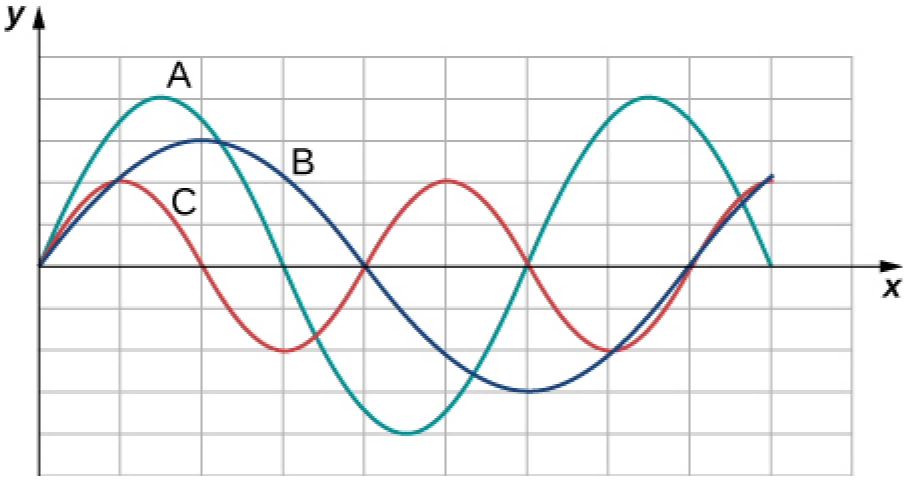
Solution:

Since the speed of a wave on a string is inversely proportional to the square root of the linear mass density, the speed would be higher in the low linear mass density of the string.

Exercise:

Problem:

Shown below are three waves that were sent down a string at different times. The tension in the string remains constant. (a) Rank the waves from the smallest wavelength to the largest wavelength. (b) Rank the waves from the lowest frequency to the highest frequency.



Exercise:

Problem:

Electrical power lines connected by two utility poles are sometimes heard to hum when driven into oscillation by the wind. The speed of the waves on the power lines depend on the tension. What provides the tension in the power lines?

Solution:

The tension in the wire is due to the weight of the electrical power cable.

Exercise:

Problem:

Two strings, one with a low mass density and one with a high linear density are spliced together. The higher density end is tied to a lab post and a student holds the free end of the low-mass density string. The student gives the string a flip and sends a pulse down the strings. If the tension is the same in both strings, does the pulse travel at the same wave velocity in both strings? If not, where does it travel faster, in the low density string or the high density string?

Problems

Exercise:

Problem:

Transverse waves are sent along a 5.00-m-long string with a speed of 30.00 m/s. The string is under a tension of 10.00 N. What is the mass of the string?

Exercise:

Problem:

A copper wire has a density of $\rho = 8920\,\text{kg/m}^3$, a radius of 1.20 mm, and a length L. The wire is held under a tension of 10.00 N. Transverse waves are sent down the wire. (a) What is the linear mass density of the wire? (b) What is the speed of the waves through the wire?

Solution:

a. $\mu = 0.040\,\text{kg/m}$; b. $v = 15.75\,\text{m/s}$

Exercise:

Problem:

A piano wire has a linear mass density of $\mu = 4.95 \ \times \ 10^{-3}\,\text{kg/m}.$ Under what tension must the string be kept to produce waves with a wave speed of 500.00 m/s?

Exercise:

Problem:

A string with a linear mass density of $\mu = 0.0060\,\text{kg/m}$ is tied to the ceiling. A 20-kg mass is tied to the free end of the string. The string is plucked, sending a pulse down the string. Estimate the speed of the pulse as it moves down the string.

Solution:

$v = 180\,\text{m/s}$

Exercise:

Problem:

A cord has a linear mass density of $\mu = 0.0075\,\text{kg/m}$ and a length of three meters. The cord is plucked and it takes 0.20 s for the pulse to reach the end of the string. What is the tension of the string?

Exercise:

Problem:

A string is 3.00 m long with a mass of 5.00 g. The string is held taut with a tension of 500.00 N applied to the string. A pulse is sent down the string. How long does it take the pulse to travel the 3.00 m of the string?

Solution:

$v = 547.723\,\text{m/s},\ \Delta t = 5.48\,\text{ms}$

Exercise:

Problem:

Two strings are attached to poles, however the first string is twice as long as the second. If both strings have the same tension and mu, what is the ratio of the speed of the pulse of the wave from the first string to the second string?

Exercise:

Problem:

Two strings are attached to poles, however the first string is twice the linear mass density mu of the second. If both strings have the same tension, what is the ratio of the speed of the pulse of the wave from the first string to the second string?

Solution:

0.707

Exercise:

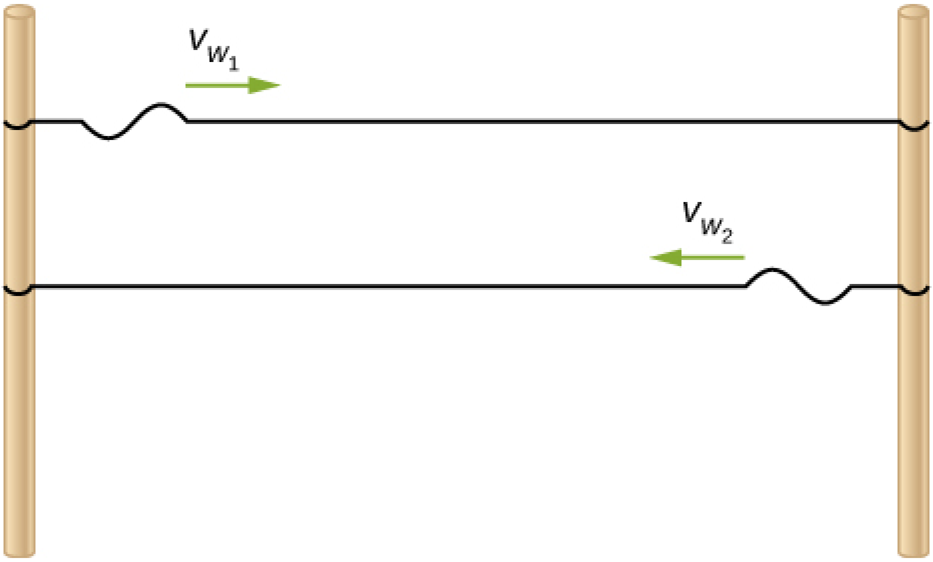
Problem:

Transverse waves travel through a string where the tension equals 7.00 N with a speed of 20.00 m/s. What tension would be required for a wave speed of 25.00 m/s?

Exercise:

Problem:

Two strings are attached between two poles separated by a distance of 2.00 m as shown below, both under the same tension of 600.00 N. String 1 has a linear density of $\mu_1 = 0.0025\,\text{kg/m}$ and string 2 has a linear mass density of $\mu_2 = 0.0035\,\text{kg/m}$. Transverse wave pulses are generated simultaneously at opposite ends of the strings. How much time passes before the pulses pass one another?



Solution:

$v_1t + v_2t = 2.00\,\text{m},\ t = 1.69\,\text{ms}$

Exercise:

Problem:

Two strings are attached between two poles separated by a distance of 2.00 meters as shown in the preceding figure, both strings have a linear density of $\mu_1 = 0.0025\,\text{kg/m}$, the tension in string 1 is 600.00 N and the tension in string 2 is 700.00 N. Transverse wave pulses are generated simultaneously at opposite ends of the strings. How much time passes before the pulses pass one another?

Exercise:

Problem:

The note E_4 is played on a piano and has a frequency of $f = 393.88$. If the linear mass density of this string of the piano is $\mu = 0.012\,\text{kg/m}$ and the string is under a tension of 1000.00 N, what is the speed of the wave on the string and the wavelength of the wave?

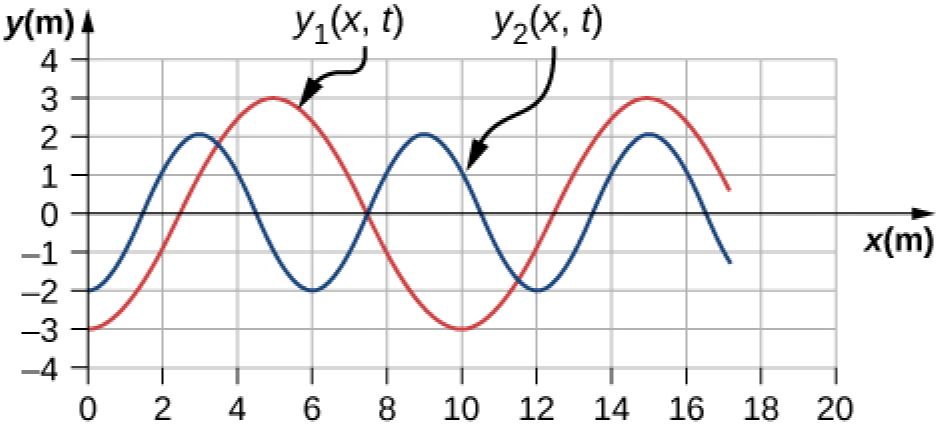
Solution:

$v=288.68\,\text{m/s},\ \lambda=0.73\,\text{m}$

Exercise:

Problem:

Two transverse waves travel through a taut string. The speed of each wave is $v=30.00\,\text{m/s}$. A plot of the vertical position as a function of the horizontal position is shown below for the time $t=0.00\,\text{s}$. (a) What is the wavelength of each wave? (b) What is the frequency of each wave? (c) What is the maximum vertical speed of each string?



Exercise:

Problem:

A sinusoidal wave travels down a taut, horizontal string with a linear mass density of $\mu=0.060\,\text{kg/m}$. The maximum vertical speed of the wave is $v_{y\,\text{max}}=0.30\,\text{cm/s}$. The wave is modeled with the wave equation $y(x,t)=A\sin\left(6.00\,\text{m}^{-1}x-24.00\,\text{s}^{-1}t\right)$. (a) What is the amplitude of the wave? (b) What is the tension in the string?

Solution:

a. $A=0.0125\,\text{cm}$; b. $F_T=0.96\,\text{N}$

Exercise:

Problem:

The speed of a transverse wave on a string is $v = 60.00\,\mathrm{m/s}$ and the tension in the string is $F_T = 100.00\,\mathrm{N}$. What must the tension be to increase the speed of the wave to $v = 120.00\,\mathrm{m/s}$?

Energy and Power of a Wave

By the end of this section, you will be able to:

- Explain how energy travels with a pulse or wave
- Describe, using a mathematical expression, how the energy in a wave depends on the amplitude of the wave

All waves carry energy, and sometimes this can be directly observed. Earthquakes can shake whole cities to the ground, performing the work of thousands of wrecking balls ([link]). Loud sounds can pulverize nerve cells in the inner ear, causing permanent hearing loss. Ultrasound is used for deep-heat treatment of muscle strains. A laser beam can burn away a malignancy. Water waves chew up beaches.

The destructive effect of an earthquake is observable evidence of the energy carried in these waves. The Richter scale rating of earthquakes is a logarithmic scale related to both their amplitude and the energy they carry.

In this section, we examine the quantitative expression of energy in waves. This will be of fundamental importance in later discussions of waves, from sound to light to quantum mechanics.

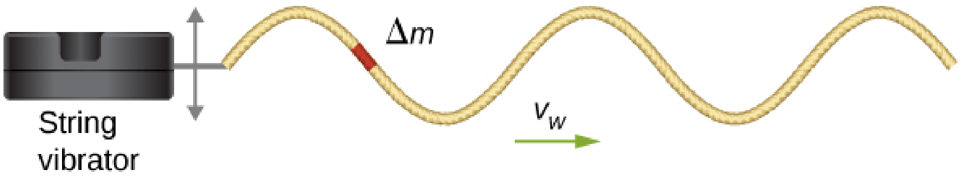
Energy in Waves

The amount of energy in a wave is related to its amplitude and its frequency. Large-amplitude earthquakes produce large ground displacements. Loud sounds have high-pressure amplitudes and come from larger-amplitude source vibrations than soft sounds. Large ocean breakers churn up the shore more than small ones. Consider the example of the seagull and the water wave earlier in the chapter ([link]). Work is done on the seagull by the wave as the seagull is moved up, changing its potential energy. The larger the amplitude, the higher the seagull is lifted by the wave and the larger the change in potential energy.

The energy of the wave depends on both the amplitude and the frequency. If the energy of each wavelength is considered to be a discrete packet of energy, a high-frequency wave will deliver more of these packets per unit time than a low-frequency wave. We will see that the average rate of energy transfer in mechanical waves is proportional to both the square of the amplitude and the square of the frequency. If two mechanical waves have equal amplitudes, but one wave has a frequency equal to twice the frequency of the other, the higher-frequency wave will have a rate of energy transfer a factor of four times as great as the rate of energy transfer of the lower-frequency wave. It should be noted that although the rate of energy transport is proportional to both the square of the amplitude and square of the frequency in mechanical waves, the rate of energy transfer in electromagnetic waves is proportional to the square of the amplitude, but independent of the frequency.

Power in Waves

Consider a sinusoidal wave on a string that is produced by a string vibrator, as shown in [link]. The string vibrator is a device that vibrates a rod up and down. A string of uniform linear mass density is attached to the rod, and the rod oscillates the string, producing a sinusoidal wave. The rod does work on the string, producing energy that propagates along the string. Consider a mass element of the string with a mass Δm, as seen in [link]. As the energy propagates along the string, each mass element of the string is driven up and down at the same frequency as the wave. Each mass element of the string can be modeled as a simple harmonic oscillator. Since the string has a constant linear density $\mu = \frac{\Delta m}{\Delta x}$, each mass element of the string has the mass $\Delta m = \mu \Delta x$.



A string vibrator is a device that vibrates a rod. A string is attached to the rod, and the rod does work on the string, driving the string up and down. This produces a sinusoidal wave in the string, which moves with a wave velocity *v*. The wave speed depends on the tension in the string and the linear mass density of the string. A section of the string with mass Δm oscillates at the same frequency as the wave.

The total mechanical energy of the wave is the sum of its kinetic energy and potential energy. The kinetic energy $K = \frac{1}{2}mv^2$ of each mass element of the string of length Δx is $\Delta K = \frac{1}{2}(\Delta m)v_y^2$, as the mass element oscillates perpendicular to the direction of the motion of the wave. Using the constant linear mass density, the kinetic energy of each mass element of the string with length Δx is

Equation:

$$\Delta K = \frac{1}{2}(\mu \Delta x)v_y^2.$$

A differential equation can be formed by letting the length of the mass element of the string approach zero,

Equation:

$$dK = \lim_{\Delta x \to 0} \frac{1}{2}(\mu \Delta x)v_y^2 = \frac{1}{2}(\mu dx)v_y^2.$$

Since the wave is a sinusoidal wave with an angular frequency ω, the position of each mass element may be modeled as $y\left(x,t\right) = A\sin\left(kx - \omega t\right)$. Each mass element of the string oscillates with a velocity $v_y = \frac{\partial y(x,t)}{\partial t} = -A\omega\cos\left(kx - \omega t\right)$. The kinetic energy of each mass element of the string becomes

Equation:

$$\begin{aligned} dK &= \tfrac{1}{2}(\mu dx)(-A\omega\cos\left(kx - \omega t\right))^2, \\ &= \tfrac{1}{2}(\mu dx)A^2\omega^2\cos^2\left(kx - \omega t\right). \end{aligned}$$

The wave can be very long, consisting of many wavelengths. To standardize the energy, consider the kinetic energy associated with a wavelength of the wave. This kinetic energy can be integrated over the wavelength to find the energy associated with each wavelength of the wave:

Equation:

$$\begin{aligned} dK &= \tfrac{1}{2}(\mu dx)A^2\omega^2\cos^2(kx), \\ \int_0^{K_\lambda} dK &= \int_0^{\lambda} \frac{1}{2}\mu A^2\omega^2\cos^2(kx)dx = \tfrac{1}{2}\mu A^2\omega^2\int_0^{\lambda}\cos^2(kx)dx, \\ K_\lambda &= \tfrac{1}{2}\mu A^2\omega^2\Big[\tfrac{1}{2}x + \tfrac{1}{4k}\sin(2kx)\Big]_0^{\lambda} = \tfrac{1}{2}\mu A^2\omega^2\left[\tfrac{1}{2}\lambda + \tfrac{1}{4k}\sin(2k\lambda) - \tfrac{1}{4k}\sin(0)\right], \\ K_\lambda &= \tfrac{1}{4}\mu A^2\omega^2\lambda. \end{aligned}$$

There is also potential energy associated with the wave. Much like the mass oscillating on a spring, there is a conservative restoring force that, when the mass element is displaced from the equilibrium position, drives the mass element back to the equilibrium position. The potential energy of the mass element can be found by considering the linear restoring force of the string, In Oscillations, we saw that the potential energy stored in a spring with a linear restoring force is equal to $U = \frac{1}{2}k_sx^2$, where the equilibrium position is defined as $x = 0.00\,\text{m}$. When a mass attached to the spring oscillates in simple harmonic motion, the angular frequency is equal to $\omega = \sqrt{\frac{k_s}{m}}$. As each mass element oscillates in simple harmonic motion, the spring constant is equal to $k_s = \Delta m\omega^2$. The potential energy of the mass element is equal to

Equation:

$$\Delta U = \frac{1}{2}k_sx^2 = \frac{1}{2}\Delta m\omega^2x^2.$$

Note that k_s is the spring constant and not the wave number $k = \frac{2\pi}{\lambda}$. This equation can be used to find the energy over a wavelength. Integrating over the wavelength, we can compute the potential energy over a wavelength:

Equation:

$$\begin{aligned} dU &= \tfrac{1}{2}k_sx^2 = \tfrac{1}{2}\mu\omega^2x^2dx, \\ U_\lambda &= \tfrac{1}{2}\mu\omega^2A^2\int_0^{\lambda}\cos^2(kx)dx = \tfrac{1}{4}\mu A^2\omega^2\lambda. \end{aligned}$$

The potential energy associated with a wavelength of the wave is equal to the kinetic energy associated with a wavelength.

The total energy associated with a wavelength is the sum of the potential energy and the kinetic energy:

Equation:

$$\begin{aligned}E_\lambda &= U_\lambda + K_\lambda,\\ E_\lambda &= \tfrac{1}{4}\mu A^2\omega^2\lambda + \tfrac{1}{4}\mu A^2\omega^2\lambda = \tfrac{1}{2}\mu A^2\omega^2\lambda.\end{aligned}$$

The time-averaged power of a sinusoidal mechanical wave, which is the average rate of energy transfer associated with a wave as it passes a point, can be found by taking the total energy associated with the wave divided by the time it takes to transfer the energy. If the velocity of the sinusoidal wave is constant, the time for one wavelength to pass by a point is equal to the period of the wave, which is also constant. For a sinusoidal mechanical wave, the time-averaged power is therefore the energy associated with a wavelength divided by the period of the wave. The wavelength of the wave divided by the period is equal to the velocity of the wave,

Note:
Equation:

$$P_{\text{ave}} = \frac{E_\lambda}{T} = \frac{1}{2}\mu A^2\omega^2\frac{\lambda}{T} = \frac{1}{2}\mu A^2\omega^2 v.$$

Note that this equation for the time-averaged power of a sinusoidal mechanical wave shows that the power is proportional to the square of the amplitude of the wave and to the square of the angular frequency of the wave. Recall that the angular frequency is equal to $\omega = 2\pi f$, so the power of a mechanical wave is equal to the square of the amplitude and the square of the frequency of the wave.

Example:
Power Supplied by a String Vibrator
Consider a two-meter-long string with a mass of 70.00 g attached to a string vibrator as illustrated in [link]. The tension in the string is 90.0 N. When the string vibrator is turned on, it oscillates with a frequency of 60 Hz and produces a sinusoidal wave on the string with an amplitude of 4.00 cm and a constant wave speed. What is the time-averaged power supplied to the wave by the string vibrator?

Strategy
The power supplied to the wave should equal the time-averaged power of the wave on the string. We know the mass of the string (m_s), the length of the string (L_s), and the tension (F_T) in the string. The speed of the wave on the string can be derived from the linear mass density and the tension. The string oscillates with the same frequency as the string vibrator, from which we can find the angular frequency.

Solution

1. Begin with the equation of the time-averaged power of a sinusoidal wave on a string:
   Equation:

   $$P = \frac{1}{2}\mu A^2\omega^2 v.$$

   The amplitude is given, so we need to calculate the linear mass density of the string, the angular frequency of the wave on the string, and the speed of the wave on the string.
2. We need to calculate the linear density to find the wave speed:
   Equation:

   $$\mu = \frac{m_s}{L_s} = \frac{0.070\,\text{kg}}{2.00\,\text{m}} = 0.035\,\text{kg/m}.$$

3. The wave speed can be found using the linear mass density and the tension of the string:
   Equation:

   $$v = \sqrt{\frac{F_T}{\mu}} = \sqrt{\frac{90.00\,\text{N}}{0.035\,\text{kg/m}}} = 50.71\,\text{m/s}.$$

4. The angular frequency can be found from the frequency:
   Equation:

   $$\omega = 2\pi f = 2\pi\left(60\,\text{s}^{-1}\right) = 376.80\,\text{s}^{-1}.$$

5. Calculate the time-averaged power:
   Equation:

   $$P = \frac{1}{2}\mu A^2\omega^2 v = \frac{1}{2}\left(0.035\frac{\text{kg}}{\text{m}}\right)(0.040\,\text{m})^2\left(376.80\,\text{s}^{-1}\right)^2\left(50.71\frac{\text{m}}{\text{s}}\right) = 201.59\,\text{W}.$$

Significance
The time-averaged power of a sinusoidal wave is proportional to the square of the amplitude of the wave and the square of the angular frequency of the wave. This is true for most mechanical waves. If either the angular frequency or the amplitude of the wave were doubled, the power would increase by a factor of four. The time-averaged power of the wave on a string is also proportional to the speed of the sinusoidal wave on the string. If the speed were doubled, by increasing the tension by a factor of four, the power would also be doubled.

Note:
Exercise:

Problem:

Check Your Understanding Is the time-averaged power of a sinusoidal wave on a string proportional to the linear density of the string?

Solution:

At first glance, the time-averaged power of a sinusoidal wave on a string may look proportional to the linear density of the string because $P = \frac{1}{2}\mu A^2\omega^2 v$; however, the speed of the wave depends on the linear density. Replacing the wave speed with $\sqrt{\frac{F_T}{\mu}}$ shows that the power is proportional to the square root of tension and proportional to the square root of the linear mass density:
$P = \frac{1}{2}\mu A^2\omega^2 v = \frac{1}{2}\mu A^2\omega^2\sqrt{\frac{F_T}{\mu}} = \frac{1}{2}A^2\omega^2\sqrt{\mu F_T}.$

The equations for the energy of the wave and the time-averaged power were derived for a sinusoidal wave on a string. In general, the energy of a mechanical wave and the power are proportional to the amplitude squared and to the angular frequency squared (and therefore the frequency squared).

Another important characteristic of waves is the intensity of the waves. Waves can also be concentrated or spread out. Waves from an earthquake, for example, spread out over a larger area as they move away from a source, so they do less damage the farther they get from the source. Changing the area the waves cover has important effects. All these pertinent factors are included in the definition of **intensity (*I*)** as power per unit area:

Note:
Equation:

$$I = \frac{P}{A},$$

where *P* is the power carried by the wave through area *A*. The definition of intensity is valid for any energy in transit, including that carried by waves. The SI unit for intensity is watts per square meter (W/m^2). Many waves are spherical waves that move out from a source as a sphere. For example, a sound speaker mounted on a post above the ground may produce sound waves that move away from the source as a spherical wave. Sound waves are discussed in more detail in the next chapter, but in general, the farther you are from the speaker, the less intense the sound you hear. As a spherical wave moves out from a source, the surface area of the wave increases as the radius increases $\left(A = 4\pi r^2\right)$. The intensity for a spherical wave is therefore

Note:
Equation:

$$I = \frac{P}{4\pi r^2}.$$

If there are no dissipative forces, the energy will remain constant as the spherical wave moves away from the source, but the intensity will decrease as the surface area increases.

In the case of the two-dimensional circular wave, the wave moves out, increasing the circumference of the wave as the radius of the circle increases. If you toss a pebble in a pond, the surface ripple moves out as a circular wave. As the ripple moves away from the source, the amplitude decreases. The energy of the wave spreads around a larger circumference and the amplitude decreases proportional to $\frac{1}{r}$, which is also the same in the case of a spherical wave, since intensity is proportional to the amplitude squared.

Summary

- The energy and power of a wave are proportional to the square of the amplitude of the wave and the square of the angular frequency of the wave.
- The time-averaged power of a sinusoidal wave on a string is found by $P_{\text{ave}} = \frac{1}{2}\mu A^2\omega^2 v$, where μ is the linear mass density of the string, A is the amplitude of the wave, ω is the angular frequency of the wave, and v is the speed of the wave.
- Intensity is defined as the power divided by the area. In a spherical wave, the area is $A = 4\pi r^2$ and the intensity is $I = \frac{P}{4\pi r^2}$. As the wave moves out from a source, the energy is conserved, but the intensity decreases as the area increases.

Conceptual Questions

Exercise:

Problem:

Consider a string with under tension with a constant linear mass density. A sinusoidal wave with an angular frequency and amplitude produced by some external driving force. If the frequency of the driving force is decreased to half of the original frequency, how is the time-averaged power of the wave affected? If the amplitude of the driving force is decreased by half, how is the time-averaged power affected? Explain your answer.

Solution:

The time averaged power is $P = \frac{E_\lambda}{T} = \frac{1}{2}\mu A^2\omega^2\frac{\lambda}{T} = \frac{1}{2}\mu A^2\omega^2 v$. If the frequency or amplitude is halved, the power decreases by a factor of 4.

Exercise:

Problem:

Circular water waves decrease in amplitude as they move away from where a rock is dropped. Explain why.

Exercise:

Problem:

In a transverse wave on a string, the motion of the string is perpendicular to the motion of the wave. If this is so, how is possible to move energy along the length of the string?

Solution:

As a portion on the string moves vertically, it exerts a force on the neighboring portion of the string, doing work on the portion and transferring the energy.

Exercise:

Problem:

The energy from the sun warms the portion of the earth facing the sun during the daylight hours. Why are the North and South Poles cold while the equator is quite warm?

Exercise:

Problem:

The intensity of a spherical waves decreases as the wave moves away from the source. If the intensity of the wave at the source is I_0, how far from the source will the intensity decrease by a factor of nine?

Solution:

The intensity of a spherical wave is $I = \frac{P}{4\pi r^2}$, if no energy is dissipated the intensity will decrease by a factor of nine at three meters.

Problems

Exercise:

Problem:

A string of length 5 m and a mass of 90 g is held under a tension of 100 N. A wave travels down the string that is modeled as $y\left(x,t\right)=0.01\,\text{m}\,\sin\left(0.40\,\text{m}^{-1}x-1170.12\,\text{s}^{-1}\right)$. What is the power over one wavelength?

Solution:

$v=74.54\,\text{m/s},\,P_{\lambda}=91.85\,\text{W}$

Exercise:

Problem:

Ultrasound of intensity $1.50\ \times\ 10^{2}\,\text{W/m}^{2}$ is produced by the rectangular head of a medical imaging device measuring 3.00 cm by 5.00 cm. What is its power output?

Exercise:

Problem:

The low-frequency speaker of a stereo set has a surface area of $A=0.05\,\text{m}^{2}$ and produces 1 W of acoustical power. (a) What is the intensity at the speaker? (b) If the speaker projects sound uniformly in all directions, at what distance from the speaker is the intensity $0.1\,\text{W/m}^{2}$?

Solution:

a. $I=20.0\,\text{W/m}^{2}$; b. $\begin{array}{l}I=\frac{P}{A},\;A=10.0\,\text{m}^{2}\\ A=4\pi r^{2},\;r=0.892\,\text{m}\end{array}$

Exercise:

Problem:

To increase the intensity of a wave by a factor of 50, by what factor should the amplitude be increased?

Exercise:

Problem:

A device called an insolation meter is used to measure the intensity of sunlight. It has an area of $100\,\text{cm}^{2}$ and registers 6.50 W. What is the intensity in W/m^{2}?

Solution:

$I=650\,\text{W/m}^{2}$

Exercise:

Problem:

Energy from the Sun arrives at the top of Earth’s atmosphere with an intensity of 1400 W/m^2. How long does it take for 1.80×10^9 J to arrive on an area of 1.00 m^2?

Exercise:

Problem:

Suppose you have a device that extracts energy from ocean breakers in direct proportion to their intensity. If the device produces 10.0 kW of power on a day when the breakers are 1.20 m high, how much will it produce when they are 0.600 m high?

Solution:

$$P \propto E \propto I \propto X^2 \Rightarrow \frac{P_2}{P_1} = \left(\frac{X_2}{X_1}\right)^2$$

$$P_2 = 2.50 \text{ kW}$$

Exercise:

Problem:

A photovoltaic array of (solar cells) is 10.0% efficient in gathering solar energy and converting it to electricity. If the average intensity of sunlight on one day is 70.00 W/m^2, what area should your array have to gather energy at the rate of 100 W? (b) What is the maximum cost of the array if it must pay for itself in two years of operation averaging 10.0 hours per day? Assume that it earns money at the rate of 9.00 cents per kilowatt-hour.

Exercise:

Problem:

A microphone receiving a pure sound tone feeds an oscilloscope, producing a wave on its screen. If the sound intensity is originally $2.00 \times 10^{-5} \text{ W/m}^2$, but is turned up until the amplitude increases by 30.0%, what is the new intensity?

Solution:

$$I \propto X^2 \Rightarrow \frac{I_1}{I_2} = \left(\frac{X_1}{X_2}\right)^2 \Rightarrow$$

$$I_2 = 3.38 \times 10^{-5} \text{ W/m}^2$$

Exercise:

Problem:

A string with a mass of 0.30 kg has a length of 4.00 m. If the tension in the string is 50.00 N, and a sinusoidal wave with an amplitude of 2.00 cm is induced on the string, what must the frequency be for an average power of 100.00 W?

Exercise:

Problem:

The power versus time for a point on a string $\left(\mu = 0.05\,\text{kg/m}\right)$ in which a sinusoidal traveling wave is induced is shown in the preceding figure. The wave is modeled with the wave equation $y\left(x, t\right) = A\sin\left(20.93\,\text{m}^{-1}x - \omega t\right)$. What is the frequency and amplitude of the wave?

Solution:

$f = 100.00\,\text{Hz},\ A = 1.10\,\text{cm}$

Exercise:

Problem:

A string is under tension F_{T1}. Energy is transmitted by a wave on the string at rate P_1 by a wave of frequency f_1. What is the ratio of the new energy transmission rate P_2 to P_1 if the tension is doubled?

Exercise:

Problem:

A 250-Hz tuning fork is struck and the intensity at the source is I_1 at a distance of one meter from the source. (a) What is the intensity at a distance of 4.00 m from the source? (b) How far from the tuning fork is the intensity a tenth of the intensity at the source?

Solution:

a. $I_2 = 0.063 I_1$; b. $\begin{array}{rcl} I_1 4\pi r_1^2 & = & I_2 4\pi r_2^2 \\ r_2 & = & 3.16\,\text{m} \end{array}$

Exercise:

Problem:

A sound speaker is rated at a voltage of $P = 120.00\,\text{V}$ and a current of $I = 10.00\,\text{A}$. Electrical power consumption is $P = IV$. To test the speaker, a signal of a sine wave is applied to the speaker. Assuming that the sound wave moves as a spherical wave and that all of the energy applied to the speaker is converted to sound energy, how far from the speaker is the intensity equal to $3.82\,\text{W/m}^2$?

Exercise:

Problem:

The energy of a ripple on a pond is proportional to the amplitude squared. If the amplitude of the ripple is 0.1 cm at a distance from the source of 6.00 meters, what was the amplitude at a distance of 2.00 meters from the source?

Solution:

$$2\pi r_1 A_1^2 = 2\pi r_2 A_2^2,\ A_1 = \left(\frac{r_2}{r_1}\right)^{1/2} A_1 = 0.17\,\mathrm{m}$$

Glossary

intensity (*I*)
power per unit area

Interference of Waves

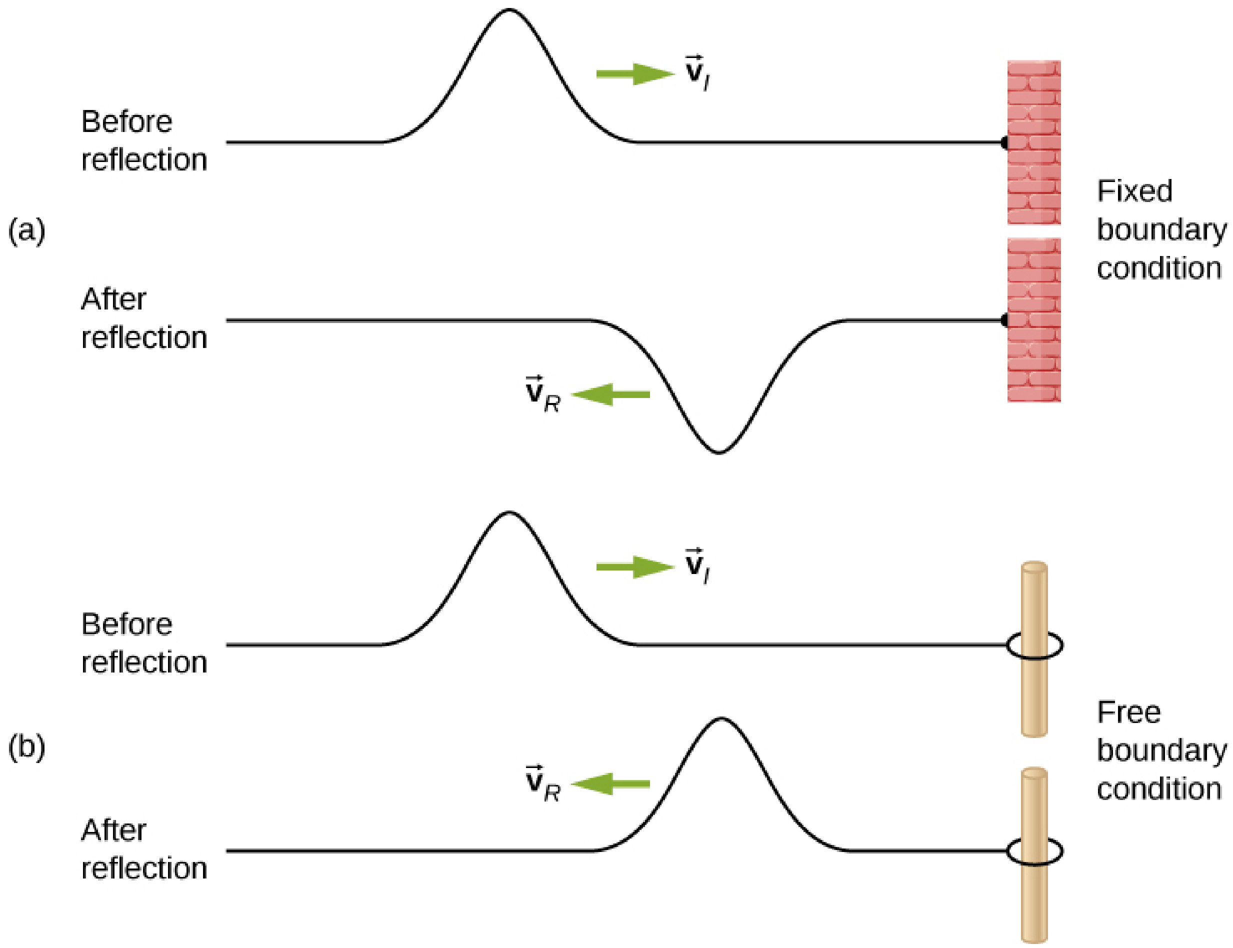
By the end of this section, you will be able to:

- Explain how mechanical waves are reflected and transmitted at the boundaries of a medium
- Define the terms interference and superposition
- Find the resultant wave of two identical sinusoidal waves that differ only by a phase shift

Up to now, we have been studying mechanical waves that propagate continuously through a medium, but we have not discussed what happens when waves encounter the boundary of the medium or what happens when a wave encounters another wave propagating through the same medium. Waves do interact with boundaries of the medium, and all or part of the wave can be reflected. For example, when you stand some distance from a rigid cliff face and yell, you can hear the sound waves reflect off the rigid surface as an echo. Waves can also interact with other waves propagating in the same medium. If you throw two rocks into a pond some distance from one another, the circular ripples that result from the two stones seem to pass through one another as they propagate out from where the stones entered the water. This phenomenon is known as interference. In this section, we examine what happens to waves encountering a boundary of a medium or another wave propagating in the same medium. We will see that their behavior is quite different from the behavior of particles and rigid bodies. Later, when we study modern physics, we will see that only at the scale of atoms do we see similarities in the properties of waves and particles.

Reflection and Transmission

When a wave propagates through a medium, it reflects when it encounters the boundary of the medium. The wave before hitting the boundary is known as the incident wave. The wave after encountering the boundary is known as the reflected wave. How the wave is reflected at the boundary of the medium depends on the boundary conditions; waves will react differently if the boundary of the medium is fixed in place or free to move ([link]). A **fixed boundary condition** exists when the medium at a boundary is fixed in place so it cannot move. A **free boundary condition** exists when the medium at the boundary is free to move.

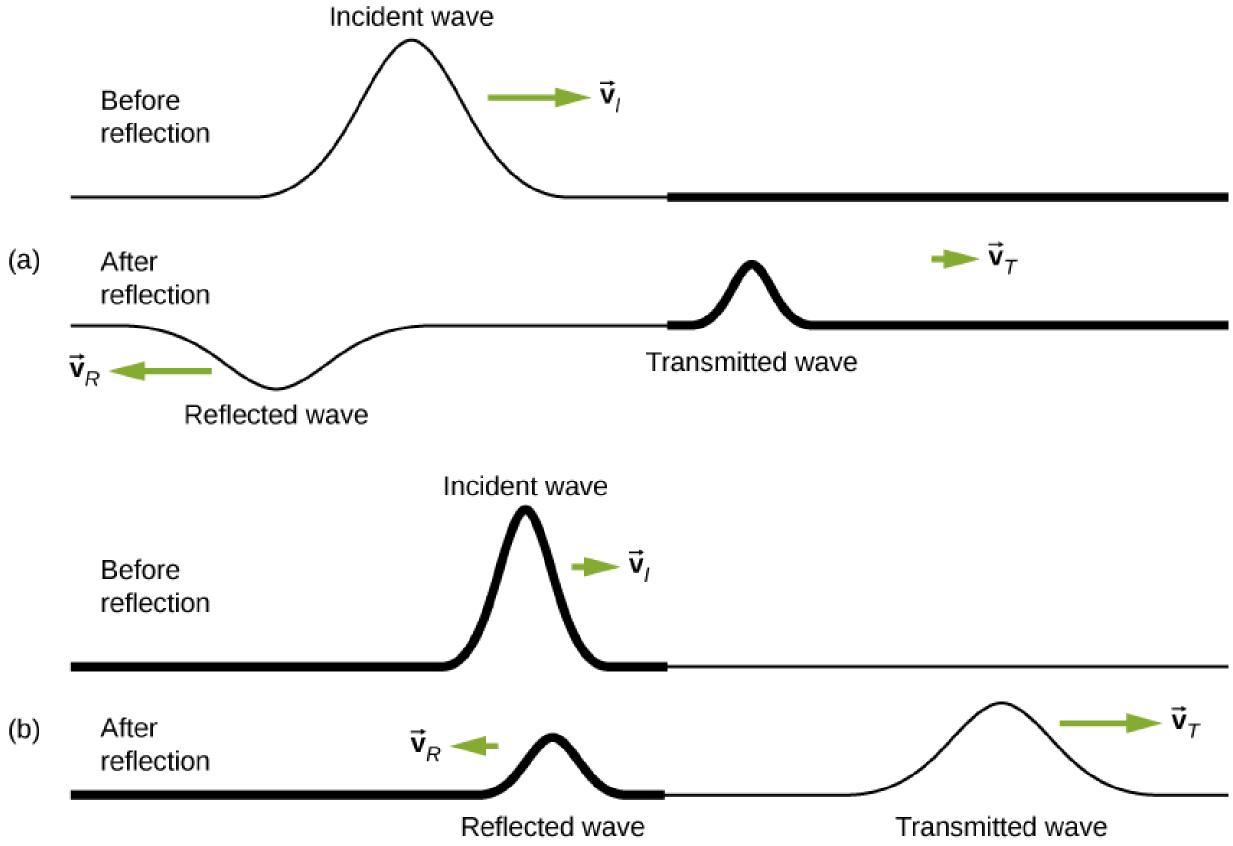

(a)
Before reflection
After reflection
$\vec{v}_I$
$\vec{v}_R$
Fixed boundary condition
(b)
Before reflection
After reflection
$\vec{v}_I$
$\vec{v}_R$
Free boundary condition

(a) One end of a string is fixed so that it cannot move. A wave propagating on the string, encountering this *fixed boundary condition*, is reflected $180^\circ (\pi \, \text{rad})$ out of phase with respect to the incident wave. (b) One end of a string is tied to a solid ring of negligible mass on a frictionless lab pole, where the ring is free to move. A wave propagating on the string, encountering this *free boundary condition*, is reflected in phase $0^\circ (0 \, \text{rad})$ with respect to the wave.

Part (a) of the [link] shows a fixed boundary condition. Here, one end of the string is fixed to a wall so the end of the string is fixed in place and the medium (the string) at the boundary cannot move. When the wave is reflected, the amplitude of the reflected way is exactly the same as the amplitude of the incident wave, but the reflected wave is reflected 180° $(\pi \, \text{rad})$ out of phase with respect to the incident wave. The phase change can be explained using Newton’s third law: Recall that Newton’s third law states that when object *A* exerts a force on object *B*, then object *B* exerts an equal and opposite force on object *A*. As the incident wave encounters the wall, the string exerts an upward force on the wall and the wall reacts by exerting an equal and opposite force on the string. The reflection at a fixed boundary is inverted. Note that the figure shows a crest of the incident wave reflected as a trough. If the incident wave were a trough, the reflected wave would be a crest.

Part (b) of the figure shows a free boundary condition. Here, one end of the string is tied to a solid ring of negligible mass on a frictionless pole, so the end of the string is free to move up and down. As the incident wave encounters the boundary of the medium, it is also reflected. In the case of a free boundary condition, the reflected wave is in phase with respect to the incident wave. In this case, the wave encounters the free boundary applying an upward force on the ring, accelerating the ring up. The ring travels up to the maximum height equal to the amplitude of the wave and then accelerates down towards the equilibrium position due to the tension in the string. The figure shows the crest of an incident wave being reflected in phase with respect to the incident wave as a crest. If the incident wave were a trough, the reflected wave would also be a trough. The amplitude of the reflected wave would be equal to the amplitude of the incident wave.

In some situations, the boundary of the medium is neither fixed nor free. Consider [link](a), where a low-linear mass density string is attached to a string of a higher linear mass density. In this case, the reflected wave is out of phase with respect to the incident wave. There is also a transmitted wave that is in phase with respect to the incident wave. Both the incident and the reflected waves have amplitudes less than the amplitude of the incident wave. If the tension is the same in both strings, the wave speed is higher in the string with the lower linear mass density.



Incident wave
Before reflection
$\vec{v}_I$
(a)
After reflection
$\vec{v}_T$
$\vec{v}_R$
Transmitted wave
Reflected wave
Incident wave
Before reflection
$\vec{v}_I$
(b)
After reflection
$\vec{v}_R$
$\vec{v}_T$
Reflected wave
Transmitted wave

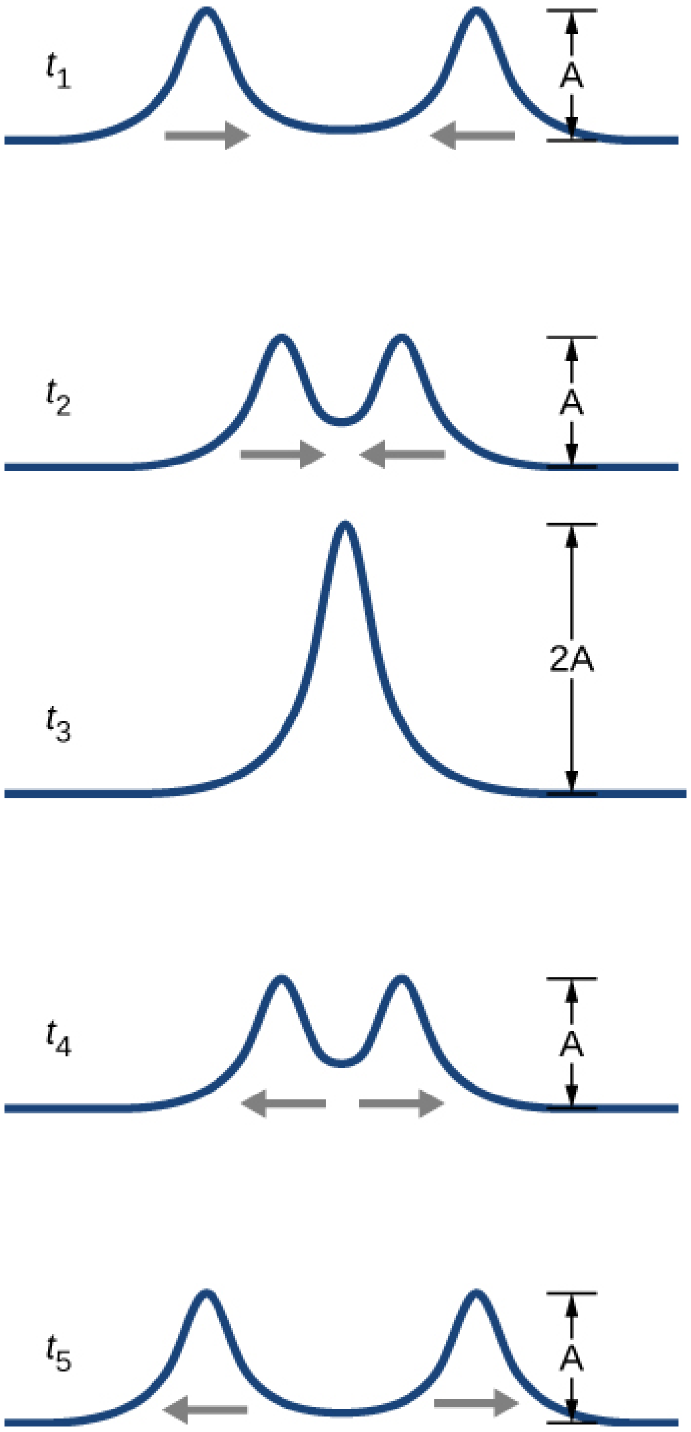
Waves traveling along two types of strings: a thick string with a high linear density and a thin string with a low linear density. Both strings are under the same tension, so a wave moves faster on the low-density string than on the high-density string. (a) A wave moving from a low-density to a high-density medium results in a reflected wave that is $180^\circ (\pi \, \text{rad})$ out of phase with respect to the incident pulse (or wave) and a transmitted wave that is in phase with the incident wave. (b) When a wave moves from a high-density medium to a low-density medium, both the reflected and transmitted wave are in phase with respect to the incident wave.

Part (b) of the figure shows a high-linear mass density string is attached to a string of a lower linear density. In this case, the reflected wave is in phase with respect to the incident wave. There is also a transmitted wave that is in phase with respect to the incident wave. Both the incident and the reflected waves have amplitudes less than the amplitude of the incident wave. Here you may notice that if the tension is the same in both strings, the wave speed is higher in the string with the lower linear mass density.

Superposition and Interference

Most waves do not look very simple. Complex waves are more interesting, even beautiful, but they look formidable. Most interesting mechanical waves consist of a combination of two or more traveling waves propagating in the same medium. The principle of superposition can be used to analyze the combination of waves.

Consider two simple pulses of the same amplitude moving toward one another in the same medium, as shown in [link]. Eventually, the waves overlap, producing a wave that has twice the amplitude, and then continue on unaffected by the encounter. The pulses are said to interfere, and this phenomenon is known as **interference**.


t_1
A
t_2
A
2A
t_3
t_4
A
t_5
A

Two pulses moving toward one another experience interference. The term interference refers to what happens when two waves overlap.

To analyze the interference of two or more waves, we use the principle of superposition. For mechanical waves, the principle of **superposition** states that if two or more traveling waves combine at the same point, the resulting position of the mass element of the medium, at that point, is the algebraic sum of the position due to the individual waves. This property is exhibited by many waves observed, such as waves on a string, sound waves, and surface water waves. Electromagnetic waves also obey the superposition principle, but the electric and magnetic fields of the combined wave are added instead of the displacement of the medium. Waves that obey the superposition principle are linear waves; waves that do not obey the superposition principle are said to be nonlinear waves. In this chapter, we deal with linear waves, in particular, sinusoidal waves.

The superposition principle can be understood by considering the linear wave equation. In Mathematics of a Wave, we defined a linear wave as a wave whose mathematical representation obeys the linear wave equation. For a transverse wave on a string with an elastic restoring force, the linear wave equation is

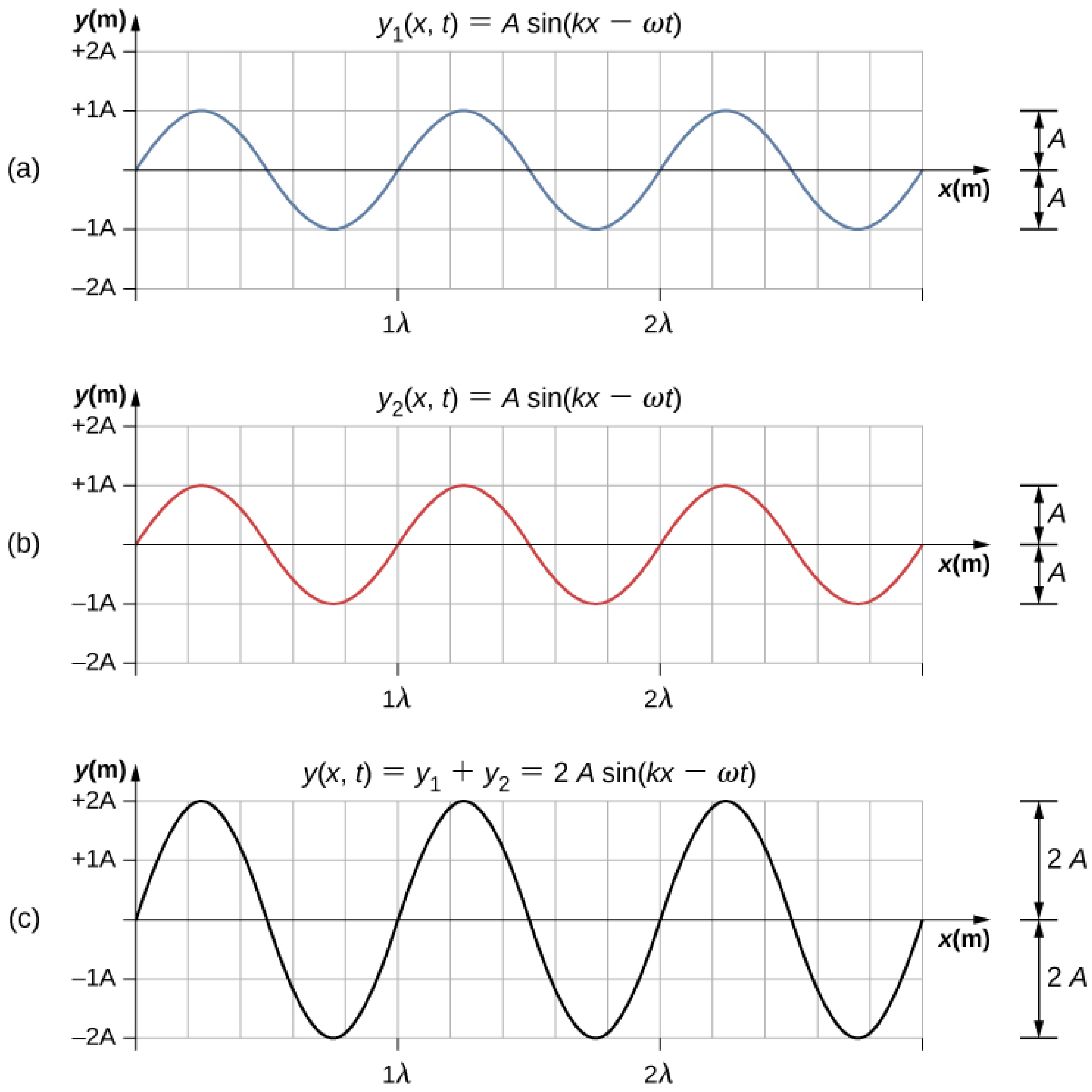
Equation:

$$\frac{\partial^2 y\left(x,t\right)}{\partial x^2}=\frac{1}{v^2}\,\frac{\partial^2 y\left(x,t\right)}{\partial t^2}.$$

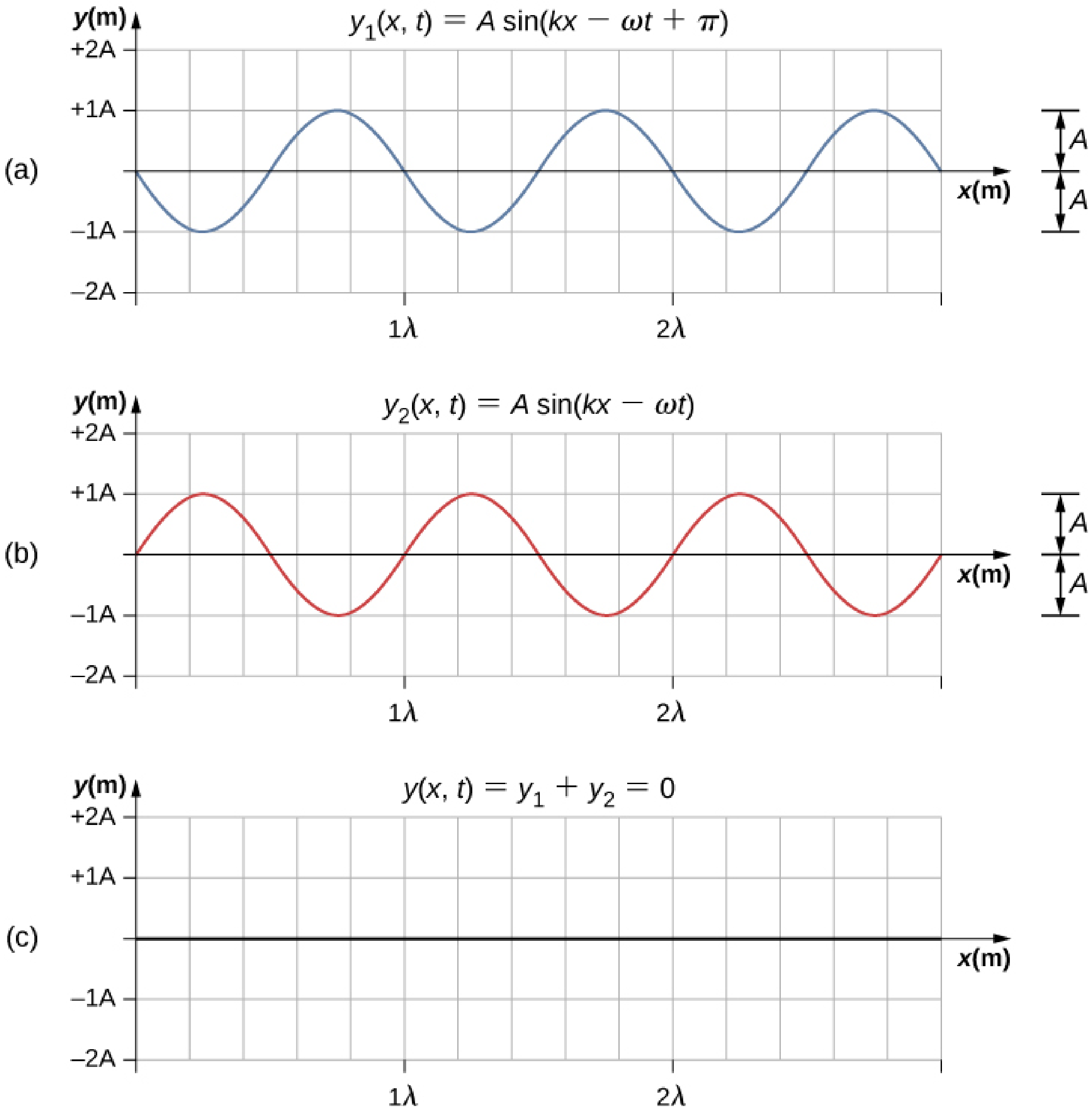
Any wave function $y(x,t)=y\left(x\mp vt\right)$, where the argument of the function is linear $\left(x\mp vt\right)$ is a solution to the linear wave equation and is a linear wave function. If wave functions $y_1\left(x,t\right)$ and $y_2\left(x,t\right)$ are solutions to the linear wave equation, the sum of the two functions $y_1\left(x,t\right)+y_2\left(x,t\right)$ is also a solution to the linear wave equation. Mechanical waves that obey superposition are normally restricted to waves with amplitudes that are small with respect to their wavelengths. If the amplitude is too large, the medium is distorted past the region where the restoring force of the medium is linear.

Waves can interfere constructively or destructively. [link] shows two identical sinusoidal waves that arrive at the same point exactly in phase. [link](a) and (b) show the two individual waves, [link](c) shows the resultant wave that results from the algebraic sum of the two linear waves. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces **constructive interference**. Because the disturbances add, constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

[link] shows two identical waves that arrive exactly 180° out of phase, producing **destructive interference**. [link](a) and (b) show the individual waves, and [link](c) shows the superposition of the two waves. Because the troughs of one wave add the crest of the other wave, the resulting amplitude is zero for destructive interference—the waves completely cancel.

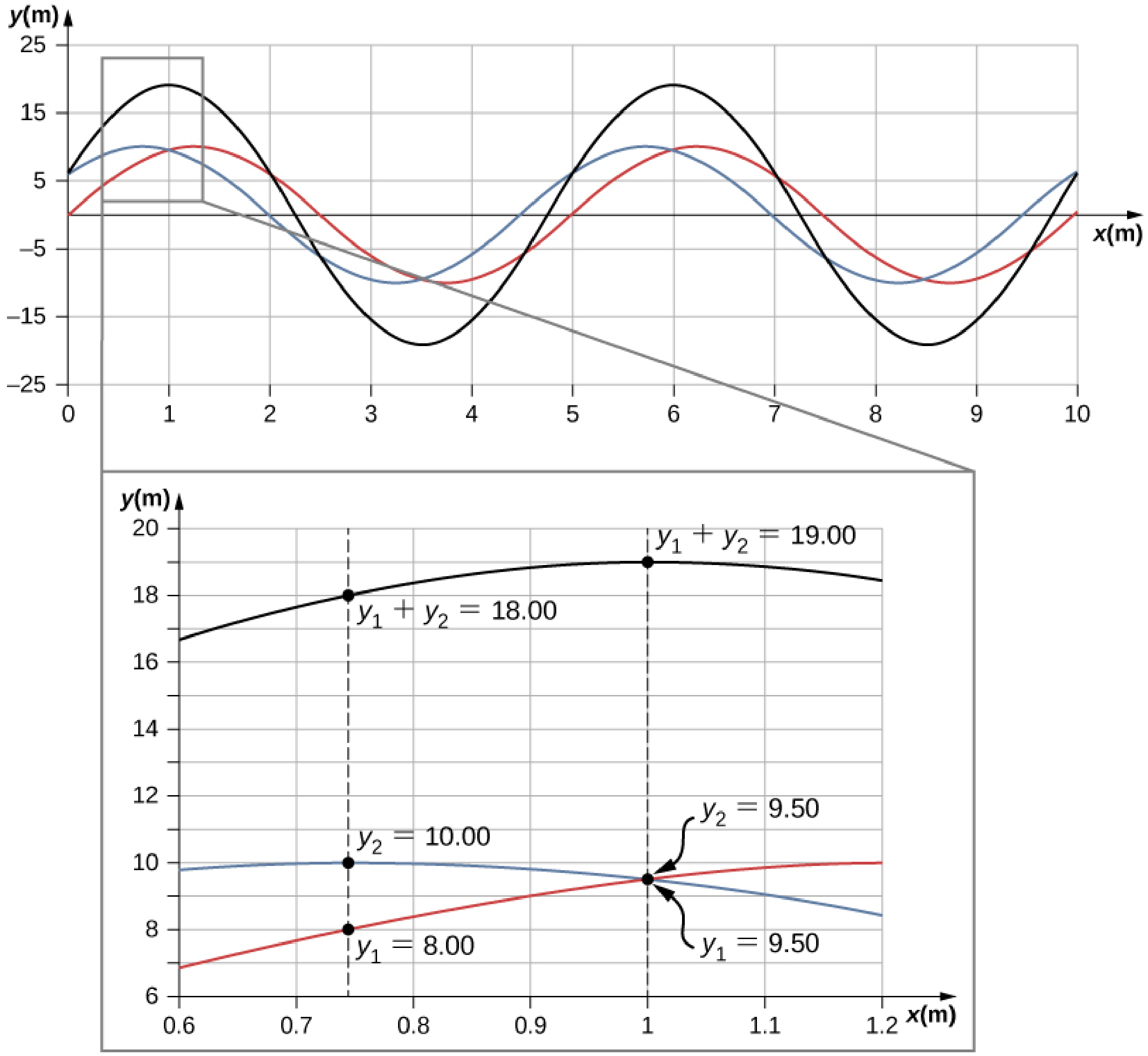


Constructive interference of two identical waves produces a wave with twice the amplitude, but the same wavelength.



Destructive interference of two identical waves, one with a phase shift of 180° (π rad), produces zero amplitude, or complete cancellation.

When linear waves interfere, the resultant wave is just the algebraic sum of the individual waves as stated in the principle of superposition. [link] shows two waves (red and blue) and the resultant wave (black). The resultant wave is the algebraic sum of the two individual waves.

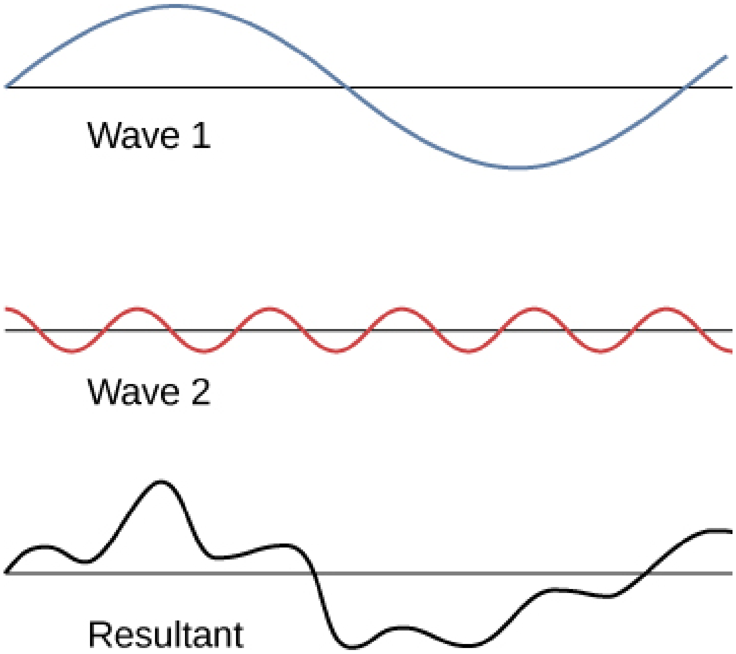


When two linear waves in the same medium interfere, the height of resulting wave is the sum of the heights of the individual waves, taken point by point. This plot shows two waves (red and blue) added together, along with the resulting wave (black). These graphs represent the height of the wave at each point. The waves may be any linear wave, including ripples on a pond, disturbances on a string, sound, or electromagnetic waves.

The superposition of most waves produces a combination of constructive and destructive interference, and can vary from place to place and time to time.

Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves interfere, and the resulting wave is the superposition of the waves.

We have shown several examples of the superposition of waves that are similar. [link] illustrates an example of the superposition of two dissimilar waves. Here again, the disturbances add, producing a resultant wave.



Superposition of nonidentical waves exhibits both constructive and destructive interference.

At times, when two or more mechanical waves interfere, the pattern produced by the resulting wave can be rich in complexity, some without any readily discernable patterns. For example, plotting the sound wave of your favorite music can look quite complex and is the superposition of the individual sound waves from many instruments; it is the complexity that makes the music interesting and worth listening to. At other times, waves can interfere and produce interesting phenomena, which are complex in their appearance and yet beautiful in simplicity of the physical principle of superposition, which formed the resulting wave. One example is the phenomenon known as standing waves, produced by two identical waves moving in different directions. We will look more closely at this phenomenon in the next section.

Note:
Try this simulation to make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern. You can observe one source or two sources. Using two sources, you can observe the interference patterns that result from varying the frequencies and the amplitudes of the sources.

Superposition of Sinusoidal Waves that Differ by a Phase Shift

Many examples in physics consist of two sinusoidal waves that are identical in amplitude, wave number, and angular frequency, but differ by a phase shift:

Equation:

$$\begin{aligned} y_1(x,t) &= A\sin(kx-\omega t+\phi), \\ y_2(x,t) &= A\sin(kx-\omega t). \end{aligned}$$

When these two waves exist in the same medium, the resultant wave resulting from the superposition of the two individual waves is the sum of the two individual waves:

Equation:

$$y_R(x,t) = y_1(x,t) + y_2(x,t) = A\sin(kx-\omega t+\phi) + A\sin(kx-\omega t).$$

The resultant wave can be better understood by using the trigonometric identity:

Equation:

$$\sin u+\sin v=2\sin\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right),$$

where $u=kx-\omega t+\phi$ and $v=kx-\omega t$. The resulting wave becomes

Equation:

$$\begin{aligned}y_R(x,t)&=y_1(x,t)+y_2(x,t)=A\sin(kx-\omega t+\phi)+A\sin(kx-\omega t)\\&=2A\sin\left(\frac{(kx-\omega t+\phi)+(kx-\omega t)}{2}\right)\cos\left(\frac{(kx-\omega t+\phi)-(kx-\omega t)}{2}\right)\\&=2A\sin\left(kx-\omega t+\frac{\phi}{2}\right)\cos\left(\frac{\phi}{2}\right).\end{aligned}$$

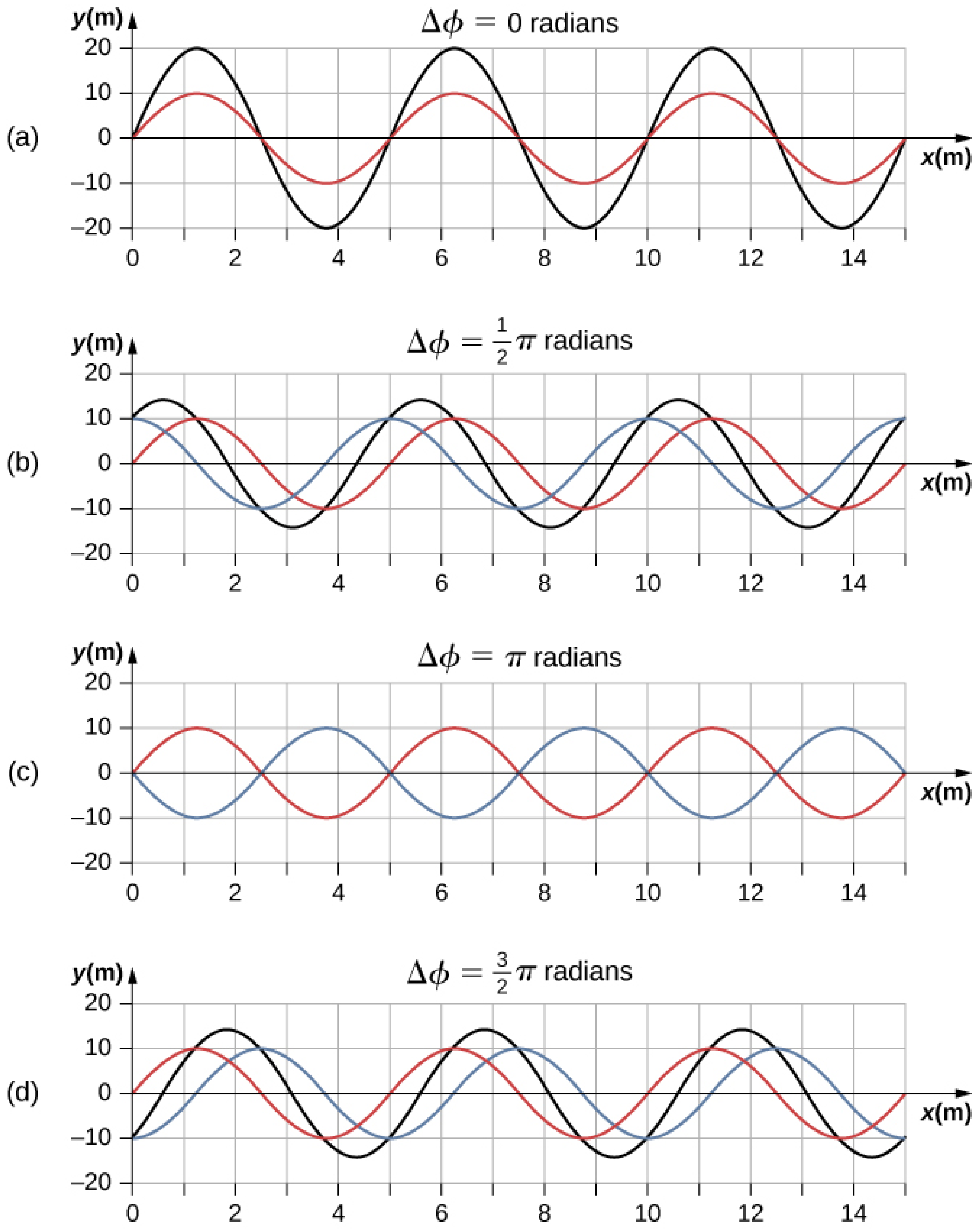
This equation is usually written as

Note:

Equation:

$$y_R(x,t)=\left[2A\cos\left(\frac{\phi}{2}\right)\right]\sin\left(kx-\omega t+\frac{\phi}{2}\right).$$

The resultant wave has the same wave number and angular frequency, an amplitude of $A_R=\left[2A\cos\left(\frac{\phi}{2}\right)\right]$, and a phase shift equal to half the original phase shift. Examples of waves that differ only in a phase shift are shown in [link]. The red and blue waves each have the same amplitude, wave number, and angular frequency, and differ only in a phase shift. They therefore have the same period, wavelength, and frequency. The green wave is the result of the superposition of the two waves. When the two waves have a phase difference of zero, the waves are in phase, and the resultant wave has the same wave number and angular frequency, and an amplitude equal to twice the individual amplitudes (part (a)). This is constructive interference. If the phase difference is 180°, the waves interfere in destructive interference (part (c)). The resultant wave has an amplitude of zero. Any other phase difference results in a wave with the same wave number and angular frequency as the two incident waves but with a phase shift of $\phi/2$ and an amplitude equal to $2A\cos(\phi/2)$. Examples are shown in parts (b) and (d).



Superposition of two waves with identical amplitudes, wavelengths, and frequency, but that differ in a phase shift. The red wave is defined by the wave function $y_1(x,t) = A\sin(kx - \omega t)$ and the blue wave is defined by the wave function $y_2(x,t) = A\sin(kx - \omega t + \phi)$. The black line shows the result of adding the two waves. The phase difference between the two waves are (a) $0.00\,\text{rad}$, (b) $\pi/2\,\text{rad}$, (c) $\pi\,\text{rad}$, and (d) $3\pi/2\,\text{rad}$.

Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs from the superposition of two identical waves that are in phase.
- Destructive interference occurs from the superposition of two identical waves that are 180° (π radians) out of phase.
- The wave that results from the superposition of two sine waves that differ only by a phase shift is a wave with an amplitude that depends on the value of the phase difference.

Conceptual Questions

Exercise:

Problem:

An incident sinusoidal wave is sent along a string that is fixed to the wall with a wave speed of *v*. The wave reflects off the end of the string. Describe the reflected wave.

Exercise:

Problem:

A string of a length of 2.00 m with a linear mass density of $\mu = 0.006\,\mathrm{kg/m}$ is attached to the end of a 2.00-m-long string with a linear mass density of $\mu = 0.012\,\mathrm{kg/m}$. The free end of the higher-density string is fixed to the wall, and a student holds the free end of the low-density string, keeping the tension constant in both strings. The student sends a pulse down the string. Describe what happens at the interface between the two strings.

Solution:

At the interface, the incident pulse produces a reflected pulse and a transmitted pulse. The reflected pulse would be out of phase with respect to the incident pulse, and would move at the same propagation speed as the incident pulse, but would move in the opposite direction. The transmitted pulse would travel in the same direction as the incident pulse, but at half the speed. The transmitted pulse would be in phase with the incident pulse. Both the reflected pulse and the transmitted pulse would have amplitudes less than the amplitude of the incident pulse.

Exercise:

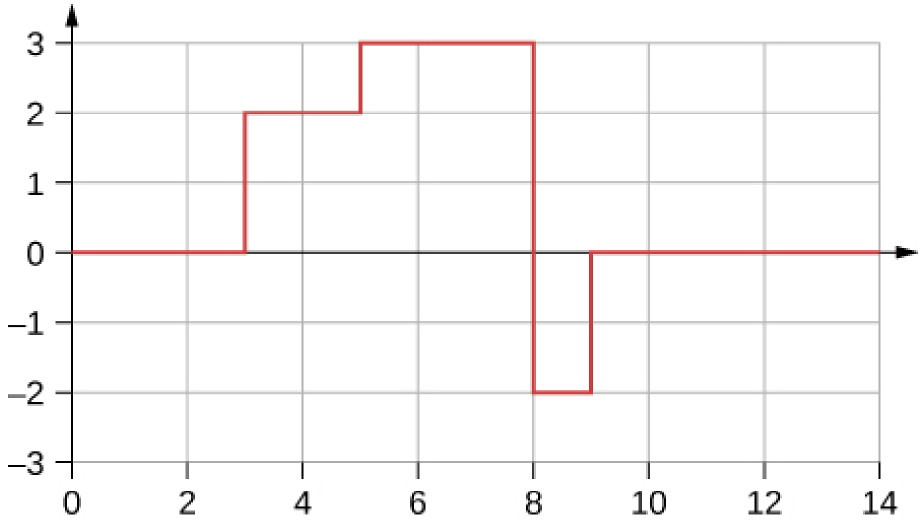
Problem:

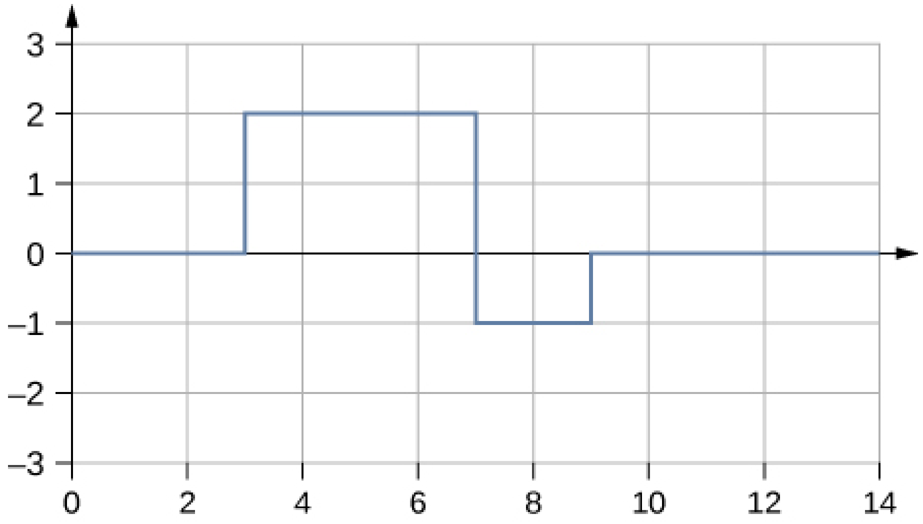
A long, tight spring is held by two students, one student holding each end. Each student gives the end a flip sending one wavelength of a sinusoidal wave down the spring in opposite directions. When the waves meet in the middle, what does the wave look like?

Exercise:

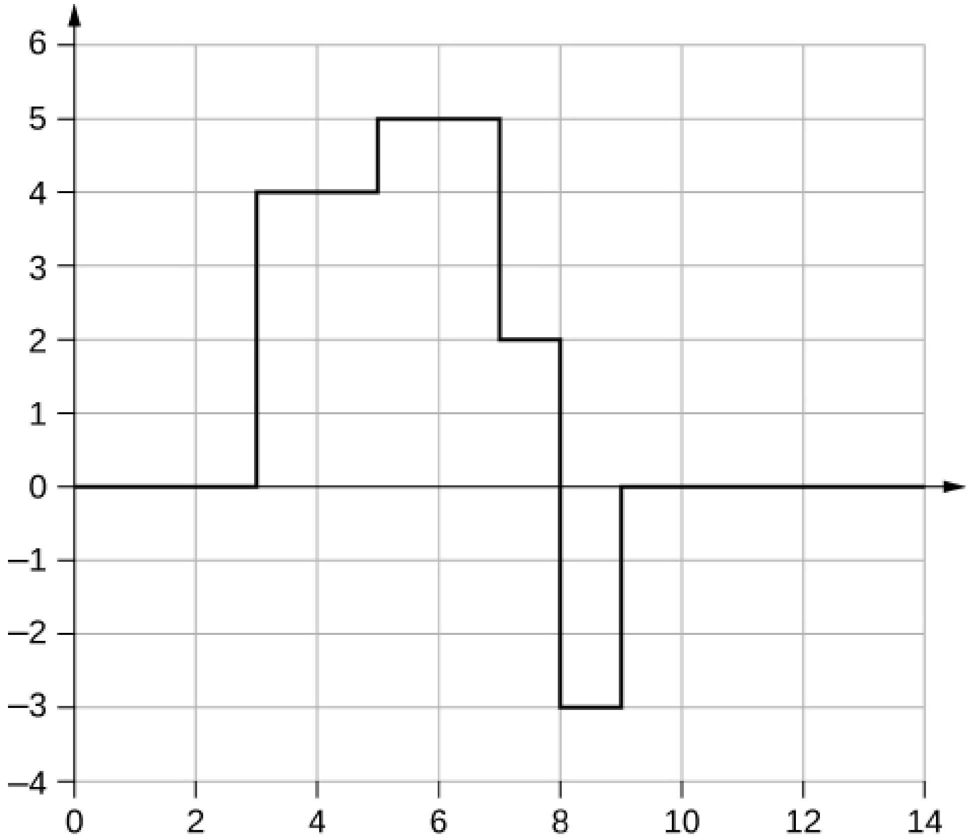
Problem:

Many of the topics discussed in this chapter are useful beyond the topics of mechanical waves. It is hard to conceive of a mechanical wave with sharp corners, but you could encounter such a wave form in your digital electronics class, as shown below. This could be a signal from a device known as an analog to digital converter, in which a continuous voltage signal is converted into a discrete signal or a digital recording of sound. What is the result of the superposition of the two signals?





Solution:



Exercise:

Problem:

A string of a constant linear mass density is held taut by two students, each holding one end. The tension in the string is constant. The students each send waves down the string by wiggling the string. (a) Is it possible for the waves to have different wave speeds? (b) Is it possible for the waves to have different frequencies? (c) Is it possible for the waves to have different wavelengths?

Problems

Exercise:

Problem:

Consider two sinusoidal waves traveling along a string, modeled as $y_1(x,t) = 0.3\,\mathrm{m}\sin\left(4\,\mathrm{m}^{-1}x + 3\,\mathrm{s}^{-1}t\right)$ and $y_2(x,t) = 0.6\,\mathrm{m}\sin\left(8\,\mathrm{m}^{-1}x - 6\,\mathrm{s}^{-1}t\right)$. What is the height of the resultant wave formed by the interference of the two waves at the position $x = 0.5\,\mathrm{m}$ at time $t = 0.2\,\mathrm{s}$?

Exercise:

Problem:

Consider two sinusoidal sine waves traveling along a string, modeled as $y_1(x,t) = 0.3\,\mathrm{m}\sin\left(4\,\mathrm{m}^{-1}x + 3\,\mathrm{s}^{-1}t + \frac{\pi}{3}\right)$ and $y_2(x,t) = 0.6\,\mathrm{m}\sin\left(8\,\mathrm{m}^{-1}x - 6\,\mathrm{s}^{-1}t\right)$. What is the height of the resultant wave formed by the interference of the two waves at the position $x = 1.0\,\mathrm{m}$ at time $t = 3.0\,\mathrm{s}$?

Solution:

$y(x,t) = 0.63\,\mathrm{m}$

Exercise:

Problem:

Consider two sinusoidal sine waves traveling along a string, modeled as $y_1(x,t) = 0.3\,\mathrm{m}\sin\left(4\,\mathrm{m}^{-1}x - 3\,\mathrm{s}^{-1}t\right)$ and $y_2(x,t) = 0.3\,\mathrm{m}\sin\left(4\,\mathrm{m}^{-1}x + 3\,\mathrm{s}^{-1}t\right)$. What is the wave function of the resulting wave? [*Hint:* Use the trig identity $\sin(u \pm v) = \sin u \cos v \pm \cos u \sin v$

Exercise:

Problem:

Two sinusoidal waves are moving through a medium in the same direction, both having amplitudes of 3.00 cm, a wavelength of 5.20 m, and a period of 6.52 s, but one has a phase shift of an angle ϕ. What is the phase shift if the resultant wave has an amplitude of 5.00 cm? [*Hint:* Use the trig identity $\sin u + \sin v = 2\sin\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right)$

Solution:

$A_R = 2A\cos\left(\frac{\phi}{2}\right),\ \phi = 1.17\,\text{rad}$

Exercise:

Problem:

Two sinusoidal waves are moving through a medium in the positive *x*-direction, both having amplitudes of 6.00 cm, a wavelength of 4.3 m, and a period of 6.00 s, but one has a phase shift of an angle $\phi = 0.50\,\text{rad}$. What is the height of the resultant wave at a time $t = 3.15\,\text{s}$ and a position $x = 0.45\,\text{m}$?

Exercise:

Problem:

Two sinusoidal waves are moving through a medium in the positive *x*-direction, both having amplitudes of 7.00 cm, a wave number of $k = 3.00\,\text{m}^{-1}$, an angular frequency of $\omega = 2.50\,\text{s}^{-1}$, and a period of 6.00 s, but one has a phase shift of an angle $\phi = \frac{\pi}{12}$ rad. What is the height of the resultant wave at a time $t = 2.00\,\text{s}$ and a position $x = 0.53\,\text{m}$?

Solution:

$y_R = 1.90\,\text{cm}$

Exercise:

Problem:

Consider two waves $y_1\left(x,t\right)$ and $y_2\left(x,t\right)$ that are identical except for a phase shift propagating in the same medium. (a)What is the phase shift, in radians, if the amplitude of the resulting wave is 1.75 times the amplitude of the individual waves? (b) What is the phase shift in degrees? (c) What is the phase shift as a percentage of the individual wavelength?

Exercise:

Problem:

Two sinusoidal waves, which are identical except for a phase shift, travel along in the same direction. The wave equation of the resultant wave is $y_R\left(x,t\right)=0.70\,\text{m}\sin\left(3.00\,\text{m}^{-1}x-6.28\,\text{s}^{-1}t+\pi/16\,\text{rad}\right)$. What are the angular frequency, wave number, amplitude, and phase shift of the individual waves?

Solution:

$$\begin{aligned}\omega &= 6.28\,\text{s}^{-1},\; k=3.00\,\text{m}^{-1},\; \phi=\tfrac{\pi}{8}\,\text{rad},\\ A_R &= 2A\cos\left(\tfrac{\phi}{2}\right),\; A=0.37\,\text{m}\end{aligned}$$

Exercise:

Problem:

Two sinusoidal waves, which are identical except for a phase shift, travel along in the same direction. The wave equation of the resultant wave is $y_R\left(x,t\right)=0.35\,\text{cm}\sin\left(6.28\,\text{m}^{-1}x-1.57\,\text{s}^{-1}t+\frac{\pi}{4}\right)$. What are the period, wavelength, amplitude, and phase shift of the individual waves?

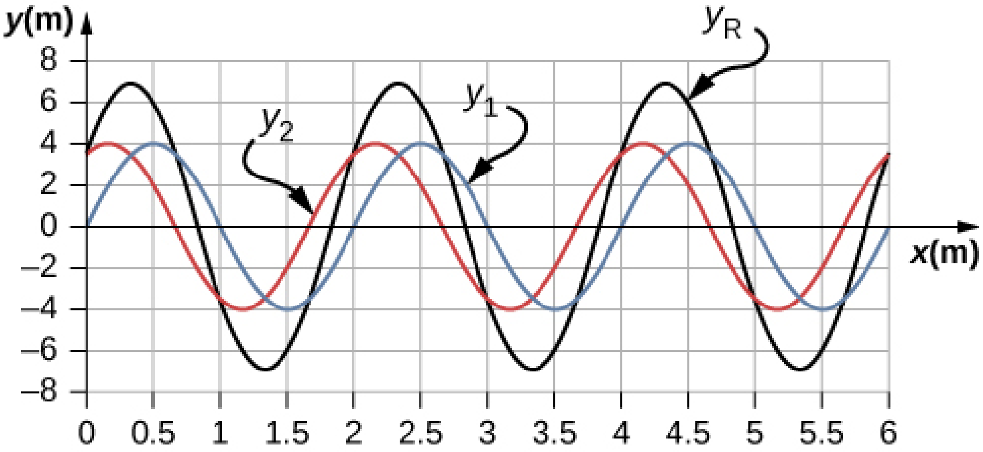
Exercise:

Problem:

Consider two wave functions, $y_1(x,t) = 4.00\,\text{m}\sin\left(\pi\,\text{m}^{-1}x - \pi\,\text{s}^{-1}t\right)$ and $y_2(x,t) = 4.00\,\text{m}\sin\left(\pi\,\text{m}^{-1}x - \pi\,\text{s}^{-1}t + \frac{\pi}{3}\right)$. (a) Using a spreadsheet, plot the two wave functions and the wave that results from the superposition of the two wave functions as a function of position $(0.00 \le x \le 6.00\,\text{m})$ for the time $t = 0.00\,\text{s}$. (b) What are the wavelength and amplitude of the two original waves? (c) What are the wavelength and amplitude of the resulting wave?

Solution:

a.



;
b. $\lambda = 2.0\,\text{m}$, $A = 4\,\text{m}$; c. $\lambda_R = 2.0\,\text{m}$, $A_R = 6.93\,\text{m}$

Exercise:

Problem:

Consider two wave functions, $y_2\left(x,t\right)=2.00\,\text{m}\sin\left(\frac{\pi}{2}\text{m}^{-1}x-\frac{\pi}{3}\text{s}^{-1}t\right)$ and $y_2\left(x,t\right)=2.00\,\text{m}\sin\left(\frac{\pi}{2}\text{m}^{-1}x-\frac{\pi}{3}\text{s}^{-1}t+\frac{\pi}{6}\right)$. (a) Verify that $y_R=2A\cos\left(\frac{\phi}{2}\right)\sin\left(kx-\omega t+\frac{\phi}{2}\right)$ is the solution for the wave that results from a superposition of the two waves. Make a column for *x*, y_1, y_2, y_1+y_2, and $y_R=2A\cos\left(\frac{\phi}{2}\right)\sin\left(kx-\omega t+\frac{\phi}{2}\right)$. Plot four waves as a function of position where the range of *x* is from 0 to 12 m.

Exercise:

Problem:

Consider two wave functions that differ only by a phase shift, $y_1\left(x,t\right)=A\cos\left(kx-\omega t\right)$ and $y_2\left(x,t\right)=A\cos\left(kx-\omega t+\phi\right)$. Use the trigonometric identities $\cos u+\cos v=2\cos\left(\frac{u-v}{2}\right)\cos\left(\frac{u+v}{2}\right)$ and $\cos\left(-\theta\right)=\cos\left(\theta\right)$ to find a wave equation for the wave resulting from the superposition of the two waves. Does the resulting wave function come as a surprise to you?

Solution:

$y_R\left(x,t\right)=2A\cos\left(\frac{\phi}{2}\right)\cos\left(kx-\omega t+\frac{\phi}{2}\right)$; The result is not surprising because $\cos\left(\theta\right)=\sin\left(\theta+\frac{\pi}{2}\right)$.

Glossary

constructive interference
: when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference
: when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fixed boundary condition
when the medium at a boundary is fixed in place so it cannot move

free boundary condition
exists when the medium at the boundary is free to move

interference
overlap of two or more waves at the same point and time

superposition
phenomenon that occurs when two or more waves arrive at the same point

Standing Waves and Resonance

By the end of this section, you will be able to:

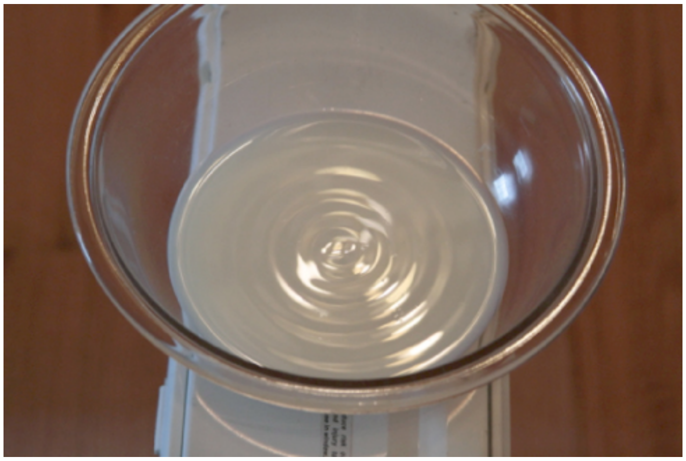
- Describe standing waves and explain how they are produced
- Describe the modes of a standing wave on a string
- Provide examples of standing waves beyond the waves on a string

Throughout this chapter, we have been studying traveling waves, or waves that transport energy from one place to another. Under certain conditions, waves can bounce back and forth through a particular region, effectively becoming stationary. These are called **standing waves**.

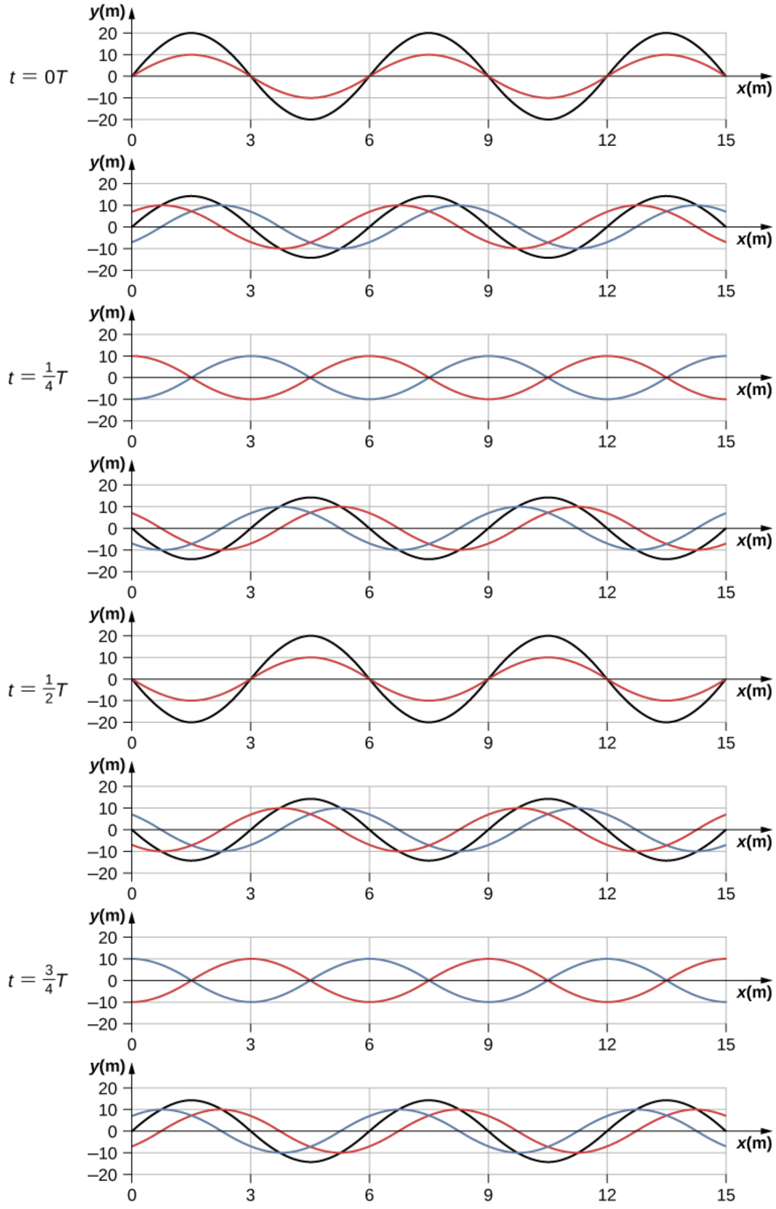
Another related effect is known as resonance. In Oscillations, we defined resonance as a phenomenon in which a small-amplitude driving force could produce large-amplitude motion. Think of a child on a swing, which can be modeled as a physical pendulum. Relatively small-amplitude pushes by a parent can produce large-amplitude swings. Sometimes this resonance is good—for example, when producing music with a stringed instrument. At other times, the effects can be devastating, such as the collapse of a building during an earthquake. In the case of standing waves, the relatively large amplitude standing waves are produced by the superposition of smaller amplitude component waves.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. You can see unmoving waves on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. [link] shows an experiment you can try at home. Take a bowl of milk and place it on a common box fan. Vibrations from the fan will produce circular standing waves in the milk. The waves are visible in the photo due to the reflection from a lamp. These waves are formed by the superposition of two or more traveling waves, such as illustrated in [link] for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a standing wave.



Standing waves are formed on the surface of a bowl of milk sitting on a box fan. The vibrations from the fan causes the surface of the milk of oscillate. The waves are visible due to the reflection of light from a lamp. (credit: David Chelton)



Time snapshots of two sine waves. The red wave is moving in the −*x*-direction and the blue wave is moving in the +*x*-direction. The resulting wave is shown in black. Consider the resultant wave at the points $x = 0\ \text{m}, 3\ \text{m}, 6\ \text{m}, 9\ \text{m}, 12\ \text{m}, 15\ \text{m}$ and notice that the resultant wave always equals zero at these points, no matter what the time is. These points are known as fixed points (nodes). In between each two nodes is an antinode, a place where the medium oscillates with an amplitude equal to the sum of the amplitudes of the individual waves.

Consider two identical waves that move in opposite directions. The first wave has a wave function of $y_1\left(x,t\right)=A\,\text{sin}\left(kx-\omega t\right)$ and the second wave has a wave function $y_2\left(x,t\right)=A\,\text{sin}\left(kx+\omega t\right)$. The waves interfere and form a resultant wave

Equation:

$$\begin{array}{l}y\left(x,t\right)=y_1\left(x,t\right)+y_2\left(x,t\right),\\ y\left(x,t\right)=A\,\text{sin}\left(kx-\omega t\right)+A\,\text{sin}\left(kx+\omega t\right).\end{array}$$

This can be simplified using the trigonometric identity

Equation:

$$\text{sin}\left(\alpha\pm\beta\right)=\text{sin}\,\alpha\,\text{cos}\,\beta\pm\text{cos}\,\alpha\,\text{sin}\,\beta,$$

where $\alpha=kx$ and $\beta=\omega t$, giving us

Equation:

$$y\left(x,t\right)=A[\text{sin}\left(kx\right)\text{cos}\left(\omega t\right)-\text{cos}\left(kx\right)\text{sin}\left(\omega t\right)+\text{sin}\left(kx\right)\text{cos}\left(\omega t\right)+\text{cos}\left(kx\right)\text{sin}\left(\omega t\right)],$$

which simplifies to

Note:

Equation:

$$y\left(x,t\right)=\left[2A\,\text{sin}\left(kx\right)\right]\text{cos}(\omega t).$$

Notice that the resultant wave is a sine wave that is a function only of position, multiplied by a cosine function that is a function only of time. Graphs of *y*(*x*,*t*) as a function of *x* for various times are shown in [link]. The red wave moves in the negative *x*-direction, the blue wave moves in the positive *x*-direction, and the black wave is the sum of the two waves. As the red and blue waves move through each other, they move in and out of constructive interference and destructive interference.

Initially, at time $t=0$, the two waves are in phase, and the result is a wave that is twice the amplitude of the individual waves. The waves are also in phase at the time $t=\frac{T}{2}$. In fact, the waves are in phase at any integer multiple of half of a period:

Equation:

$$t=n\frac{T}{2}\,\text{where}\,n=0,1,2,3....\;\text{(in phase)}.$$

At other times, the two waves are $180^\circ\,(\pi\,\text{radians})$ out of phase, and the resulting wave is equal to zero. This happens at

Equation:

$$t=\frac{1}{4}T,\,\frac{3}{4}T,\,\frac{5}{4}T,\ldots,\,\frac{n}{4}T\text{ where }n=1,3,5\ldots.\ \text{(out of phase)}.$$

Notice that some *x*-positions of the resultant wave are always zero no matter what the phase relationship is. These positions are called **nodes**. Where do the nodes occur? Consider the solution to the sum of the two waves

Equation:

$$y\left(x,t\right)=\left[2A\,\text{sin}\left(kx\right)\right]\text{cos}\left(\omega t\right).$$

Finding the positions where the sine function equals zero provides the positions of the nodes.

Equation:

$$\begin{array}{rcl}\text{sin}\left(kx\right) & = & 0\\ kx & = & 0,\pi,2\pi,3\pi,\ldots\\ \frac{2\pi}{\lambda}x & = & 0,\pi,2\pi,3\pi,\ldots\\ x & = & 0,\frac{\lambda}{2},\lambda,\frac{3\lambda}{2},\ldots=n\frac{\lambda}{2}\quad n=0,1,2,3,\ldots.\end{array}$$

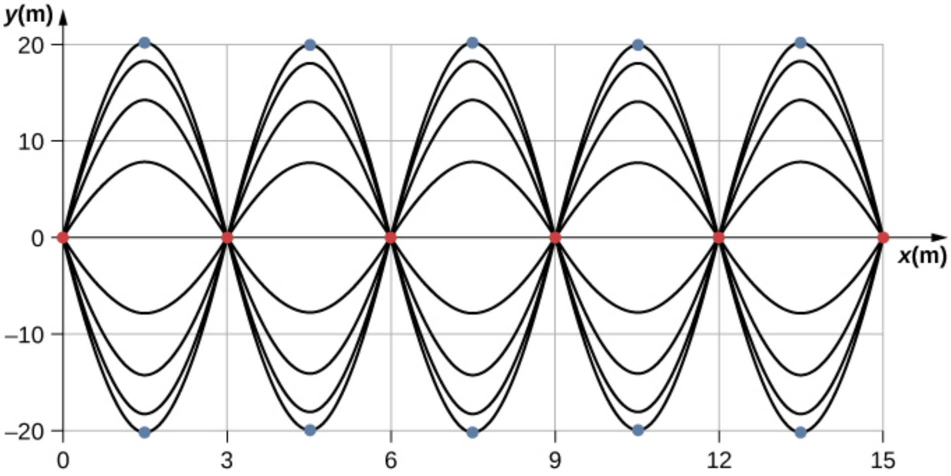
There are also positions where *y* oscillates between $y=\pm A$. These are the **antinodes**. We can find them by considering which values of *x* result in $\text{sin}\left(kx\right)=\pm 1$.

Equation:

$$\begin{array}{rcl}\text{sin}\left(kx\right) & = & \pm 1\\ kx & = & \frac{\pi}{2},\frac{3\pi}{2},\frac{5\pi}{2},\ldots\\ \frac{2\pi}{\lambda}x & = & \frac{\pi}{2},\frac{3\pi}{2},\frac{5\pi}{2},\ldots\\ x & = & \frac{\lambda}{4},\frac{3\lambda}{4},\frac{5\lambda}{4},\ldots=n\frac{\lambda}{4}\quad n=1,3,5,\ldots.\end{array}$$

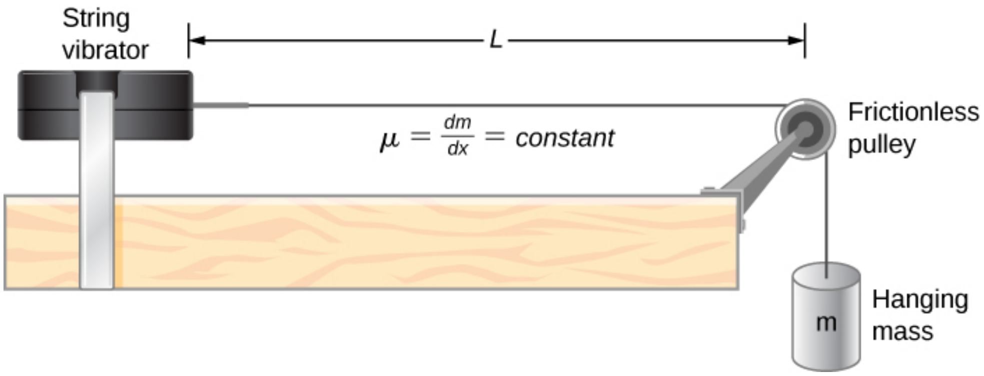
What results is a standing wave as shown in [link], which shows snapshots of the resulting wave of two identical waves moving in opposite directions. The resulting wave appears to be a sine wave with nodes at integer multiples of half wavelengths. The antinodes oscillate between $y=\pm 2A$ due to the cosine term, $\text{cos}\left(\omega t\right)$, which oscillates between ± 1.

The resultant wave appears to be standing still, with no apparent movement in the *x*-direction, although it is composed of one wave function moving in the positive, whereas the second wave is moving in the negative *x*-direction. [link] shows various snapshots of the resulting wave. The nodes are marked with red dots while the antinodes are marked with blue dots.



When two identical waves are moving in opposite directions, the resultant wave is a standing wave. Nodes appear at integer multiples of half wavelengths. Antinodes appear at odd multiples of quarter wavelengths, where they oscillate between $y = \pm A$. The nodes are marked with red dots and the antinodes are marked with blue dots.

A common example of standing waves are the waves produced by stringed musical instruments. When the string is plucked, pulses travel along the string in opposite directions. The ends of the strings are fixed in place, so nodes appear at the ends of the strings—the boundary conditions of the system, regulating the resonant frequencies in the strings. The resonance produced on a string instrument can be modeled in a physics lab using the apparatus shown in [link].



A lab setup for creating standing waves on a string. The string has a node on each end and a constant linear density. The length between the fixed boundary conditions is L. The hanging mass provides the tension in the string, and the speed of the waves on the string is proportional to the square root of the tension divided by the linear mass density.

The lab setup shows a string attached to a string vibrator, which oscillates the string with an adjustable frequency *f*. The other end of the string passes over a frictionless pulley and is tied to a hanging mass. The magnitude of the tension in the string is equal to the weight of the hanging mass. The string has a constant linear density (mass per length) μ and the speed at which a wave travels down the string equals $v=\sqrt{\frac{F_T}{\mu}}=\sqrt{\frac{mg}{\mu}}$ [link]. The symmetrical boundary conditions (a node at each end) dictate the possible frequencies that can excite standing waves. Starting from a frequency of zero and slowly increasing the frequency, the first mode $n=1$ appears as shown in [link]. The first mode, also called the fundamental mode or the first harmonic, shows half of a wavelength has formed, so the wavelength is equal to twice the length between the nodes $\lambda_1=2L$. The **fundamental frequency**, or first harmonic frequency, that drives this mode is

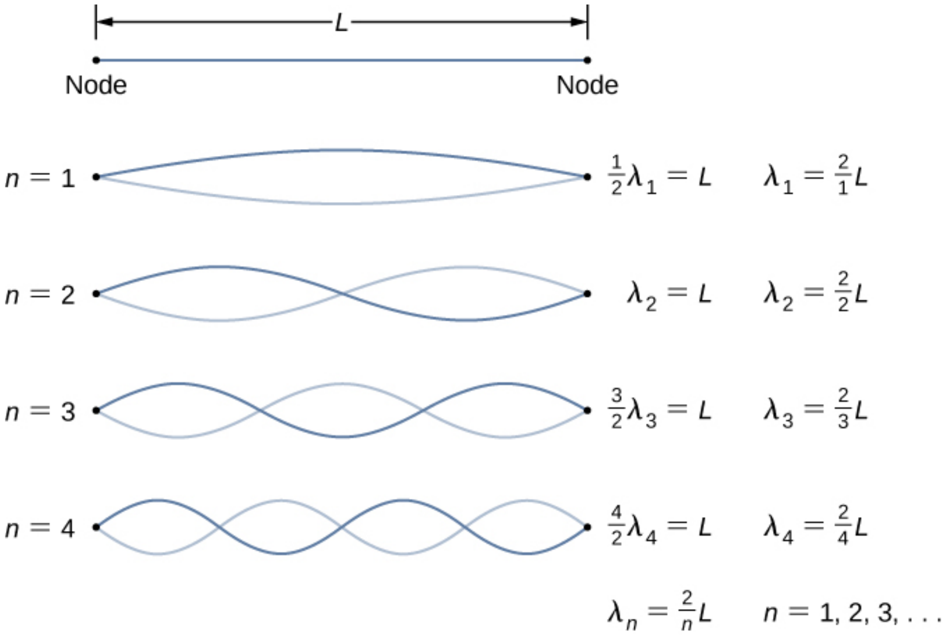
Equation:

$$f_1=\frac{v}{\lambda_1}=\frac{v}{2L},$$

where the speed of the wave is $v=\sqrt{\frac{F_T}{\mu}}$. Keeping the tension constant and increasing the frequency leads to the second harmonic or the $n=2$ mode. This mode is a full wavelength $\lambda_2=L$ and the frequency is twice the fundamental frequency:

Equation:

$$f_2=\frac{v}{\lambda_2}=\frac{v}{L}=2f_1.$$


L
Node
Node
$n = 1$
$\frac{1}{2}\lambda_1 = L$
$\lambda_1 = \frac{2}{1}L$
$n = 2$
$\lambda_2 = L$
$\lambda_2 = \frac{2}{2}L$
$n = 3$
$\frac{3}{2}\lambda_3 = L$
$\lambda_3 = \frac{2}{3}L$
$n = 4$
$\frac{4}{2}\lambda_4 = L$
$\lambda_4 = \frac{2}{4}L$
$\lambda_n = \frac{2}{n}L$
$n = 1, 2, 3, \ldots$

Standing waves created on a string of length *L*. A node occurs at each end of the string. The nodes are boundary conditions that limit the possible frequencies that excite standing waves. (Note that the amplitudes of the oscillations have been kept constant for visualization. The standing wave patterns possible on the string are known as the normal modes. Conducting this experiment in the lab would result in a decrease in amplitude as the frequency increases.)

The next two modes, or the third and fourth harmonics, have wavelengths of $\lambda_3 = \frac{2}{3}L$ and $\lambda_4 = \frac{2}{4}L$, driven by frequencies of $f_3 = \frac{3v}{2L} = 3f_1$ and $f_4 = \frac{4v}{2L} = 4f_1$. All frequencies above the frequency f_1 are known as the **overtones**. The equations for the wavelength and the frequency can be summarized as:

Note:
Equation:

$$\lambda_n = \frac{2}{n}L \quad n = 1, 2, 3, 4, 5...$$

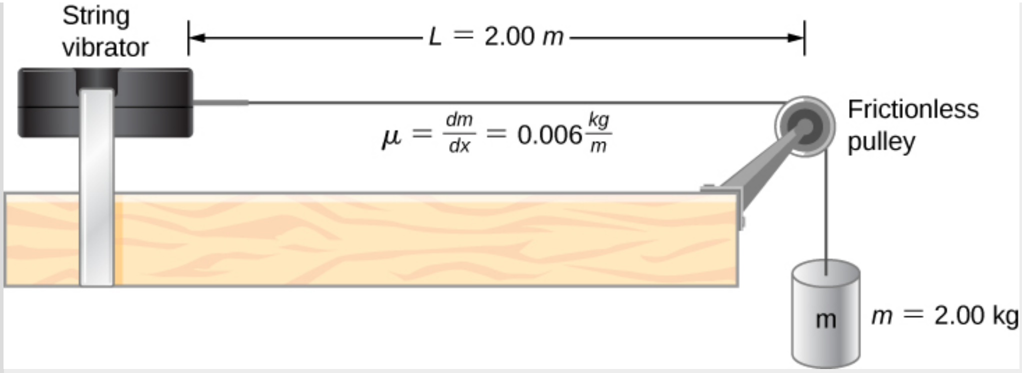
Note:
Equation:

$$f_n = n\frac{v}{2L} = nf_1 \quad n = 1, 2, 3, 4, 5...$$

The standing wave patterns that are possible for a string, the first four of which are shown in [link], are known as the **normal modes**, with frequencies known as the normal frequencies. In summary, the first frequency to produce a normal mode is called the fundamental frequency (or first harmonic). Any frequencies above the fundamental frequency are overtones. The second frequency of the $n = 2$ normal mode of the string is the first overtone (or second harmonic). The frequency of the $n = 3$ normal mode is the second overtone (or third harmonic) and so on.

The solutions shown as [link] and [link] are for a string with the boundary condition of a node on each end. When the boundary condition on either side is the same, the system is said to have symmetric boundary conditions. [link] and [link] are good for any symmetric boundary conditions, that is, nodes at both ends or antinodes at both ends.

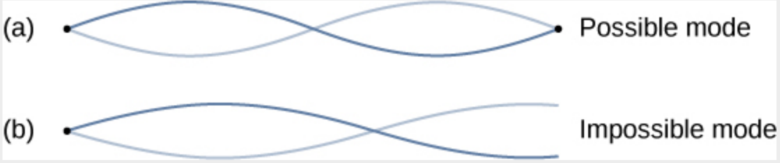
Example:
Standing Waves on a String
Consider a string of $L = 2.00 \text{ m}$. attached to an adjustable-frequency string vibrator as shown in [link]. The waves produced by the vibrator travel down the string and are reflected by the fixed boundary condition at the pulley. The string, which has a linear mass density of $\mu = 0.006 \text{ kg/m}$, is passed over a frictionless pulley of a negligible mass, and the tension is provided by a 2.00-kg hanging mass. (a) What is the velocity of the waves on the string? (b) Draw a sketch of the first three normal modes of the standing waves that can be produced on the string and label each with the wavelength. (c) List the frequencies that the string vibrator must be tuned to in order to produce the first three normal modes of the standing waves.



A string attached to an adjustable-frequency string vibrator.

Strategy

a. The velocity of the wave can be found using $v=\sqrt{\frac{F_T}{\mu}}$. The tension is provided by the weight of the hanging mass.
b. The standing waves will depend on the boundary conditions. There must be a node at each end. The first mode will be one half of a wave. The second can be found by adding a half wavelength. That is the shortest length that will result in a node at the boundaries. For example, adding one quarter of a wavelength will result in an antinode at the boundary and is not a mode which would satisfy the boundary conditions. This is shown in [link].
c. Since the wave speed velocity is the wavelength times the frequency, the frequency is wave speed divided by the wavelength.



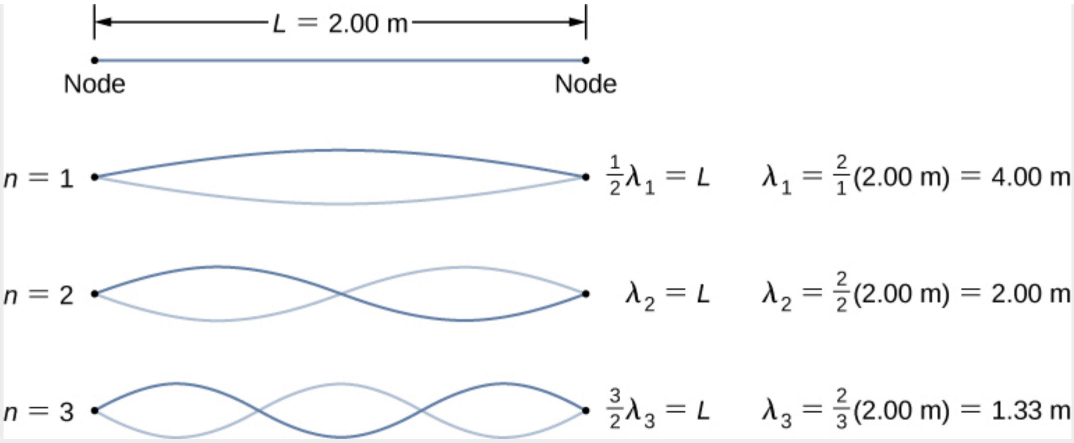
(a) The figure represents the second mode of the string that satisfies the boundary conditions of a node at each end of the string. (b)This figure could not possibly be a normal mode on the string because it does not satisfy the boundary conditions. There is a node on one end, but an antinode on the other.

Solution

a. Begin with the velocity of a wave on a string. The tension is equal to the weight of the hanging mass. The linear mass density and mass of the hanging mass are given:
Equation:

$$v=\sqrt{\frac{F_T}{\mu}}=\sqrt{\frac{mg}{\mu}}=\sqrt{\frac{2\text{ kg}\left(9.8\frac{\text{m}}{\text{s}}\right)}{0.006\frac{\text{kg}}{\text{m}}}}=57.15\text{ m/s}.$$

b. The first normal mode that has a node on each end is a half wavelength. The next two modes are found by adding a half of a wavelength.


L = 2.00 m
Node
Node
n = 1
$\frac{1}{2}\lambda_1 = L$
$\lambda_1 = \frac{2}{1}(2.00\text{ m}) = 4.00\text{ m}$
n = 2
$\lambda_2 = L$
$\lambda_2 = \frac{2}{2}(2.00\text{ m}) = 2.00\text{ m}$
n = 3
$\frac{3}{2}\lambda_3 = L$
$\lambda_3 = \frac{2}{3}(2.00\text{ m}) = 1.33\text{ m}$

c. The frequencies of the first three modes are found by using $f = \frac{v_w}{\lambda}$.

Equation:

$$f_1 = \frac{v_w}{\lambda_1} = \frac{57.15\text{ m/s}}{4.00\text{ m}} = 14.29\text{ Hz}$$
$$f_2 = \frac{v_w}{\lambda_2} = \frac{57.15\text{ m/s}}{2.00\text{ m}} = 28.58\text{ Hz}$$
$$f_3 = \frac{v_w}{\lambda_3} = \frac{57.15\text{ m/s}}{1.333\text{ m}} = 42.87\text{ Hz}$$

Significance

The three standing modes in this example were produced by maintaining the tension in the string and adjusting the driving frequency. Keeping the tension in the string constant results in a constant velocity. The same modes could have been produced by keeping the frequency constant and adjusting the speed of the wave in the string (by changing the hanging mass.)

Note:

Visit this simulation to play with a 1D or 2D system of coupled mass-spring oscillators. Vary the number of masses, set the initial conditions, and watch the system evolve. See the spectrum of normal modes for arbitrary motion. See longitudinal or transverse modes in the 1D system.

Note:

Exercise:

Problem:

Check Your Understanding The equations for the wavelengths and the frequencies of the modes of a wave produced on a string:

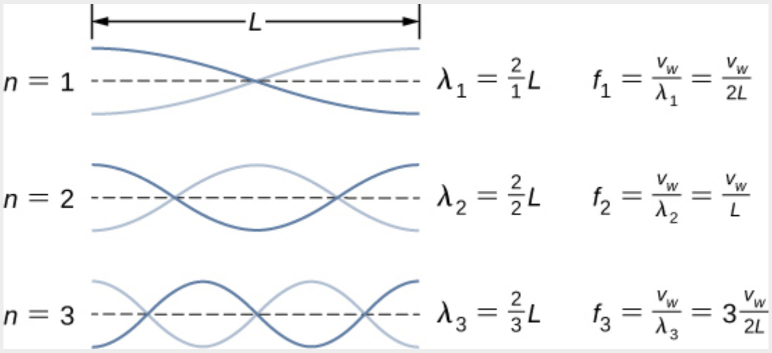
Equation:

$$\lambda_n = \frac{2}{n}L \quad n = 1, 2, 3, 4, 5\ldots \text{ and}$$
$$f_n = n\frac{v}{2L} = nf_1 \quad n = 1, 2, 3, 4, 5\ldots$$

were derived by considering a wave on a string where there were symmetric boundary conditions of a node at each end. These modes resulted from two sinusoidal waves with identical characteristics except they were moving in opposite directions, confined to a region *L* with nodes required at both ends. Will the same equations work if there were symmetric boundary conditions with antinodes at each end? What would the normal modes look like for a medium that was free to oscillate on each end? Don't worry for now if you cannot imagine such a medium, just consider two sinusoidal wave functions in a region of length *L*, with antinodes on each end.

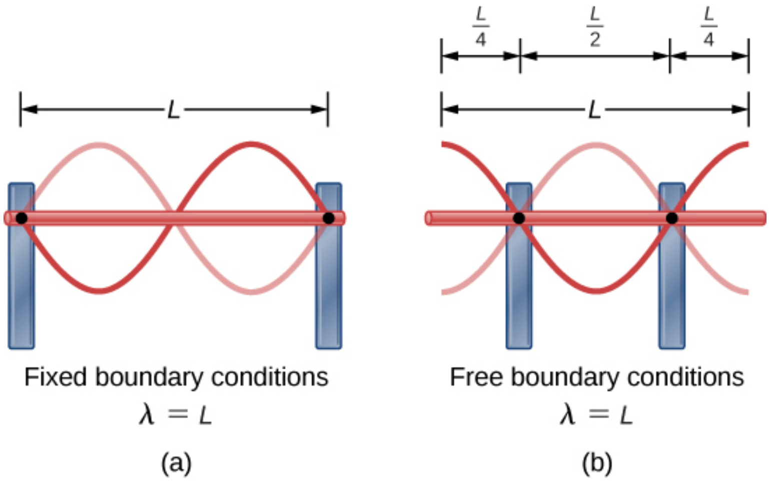
Solution:

Yes, the equations would work equally well for symmetric boundary conditions of a medium free to oscillate on each end where there was an antinode on each end. The normal modes of the first three modes are shown below. The dotted line shows the equilibrium position of the medium.

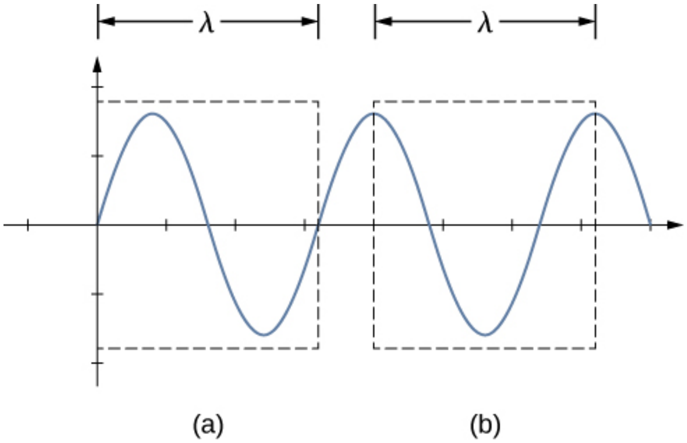

L
$n = 1$
$\lambda_1 = \frac{2}{1}L$
$f_1 = \frac{v_w}{\lambda_1} = \frac{v_w}{2L}$
$n = 2$
$\lambda_2 = \frac{2}{2}L$
$f_2 = \frac{v_w}{\lambda_2} = \frac{v_w}{L}$
$n = 3$
$\lambda_3 = \frac{2}{3}L$
$f_3 = \frac{v_w}{\lambda_3} = 3\frac{v_w}{2L}$

Note that the first mode is two quarters, or one half, of a wavelength. The second mode is one quarter of a wavelength, followed by one half of a wavelength, followed by one quarter of a wavelength, or one full wavelength. The third mode is one and a half wavelengths. These are the same result as the string with a node on each end. The equations for symmetrical boundary conditions work equally well for fixed boundary conditions and free boundary conditions. These results will be revisited in the next chapter when discussing sound wave in an open tube.

The free boundary conditions shown in the last Check Your Understanding may seem hard to visualize. How can there be a system that is free to oscillate on each end? In [link] are shown two possible configuration of a metallic rods (shown in red) attached to two supports (shown in blue). In part (a), the rod is supported at the ends, and there are fixed boundary conditions at both ends. Given the proper frequency, the rod can be driven into resonance with a wavelength equal to length of the rod, with nodes at each end. In part (b), the rod is supported at positions one quarter of the length from each end of the rod, and there are free boundary conditions at both ends. Given the proper frequency, this rod can also be driven into resonance with a wavelength equal to the length of the rod, but there are antinodes at each end. If you are having trouble visualizing the wavelength in this figure, remember that the wavelength may be measured between any two nearest identical points and consider [link].



(a) A metallic rod of length *L* (red) supported by two supports (blue) on each end. When driven at the proper frequency, the rod can resonate with a wavelength equal to the length of the rod with a node on each end. (b) The same metallic rod of length *L* (red) supported by two supports (blue) at a position a quarter of the length of the rod from each end. When driven at the proper frequency, the rod can resonate with a wavelength equal to the length of the rod with an antinode on each end.



A wavelength may be measure between the nearest two repeating points. On the wave on a string, this means the same height and slope. (a) The wavelength is measured between the two nearest points where the height is zero and the slope is maximum and positive. (b) The wavelength is measured between two identical points where the height is maximum and the slope is zero.

Note that the study of standing waves can become quite complex. In [link](a), the $n = 2$ mode of the standing wave is shown, and it results in a wavelength equal to L. In this configuration, the $n = 1$ mode would also have been possible with a standing wave equal to $2L$. Is it possible to get the $n = 1$ mode for the configuration shown in part (b)? The answer is no. In this configuration, there are additional conditions set beyond the boundary conditions. Since the rod is mounted at a point one quarter of the length from each side, a node must exist there, and this limits the possible modes of standing waves that can be created. We leave it as an exercise for the reader to consider if other modes of standing waves are possible. It should be noted that when a system is driven at a frequency that does not cause the system to resonate, vibrations may still occur, but the amplitude of the vibrations will be much smaller than the amplitude at resonance.

A field of mechanical engineering uses the sound produced by the vibrating parts of complex mechanical systems to troubleshoot problems with the systems. Suppose a part in an automobile is resonating at the frequency of the car’s engine, causing unwanted vibrations in the automobile. This may cause the engine to fail prematurely. The engineers use microphones to record the sound produced by the engine, then use a technique called Fourier analysis to find frequencies of sound produced with large amplitudes and then look at the parts list of the automobile to find a part that would resonate at that frequency. The solution may be as simple as changing the composition of the material used or changing the length of the part in question.

There are other numerous examples of resonance in standing waves in the physical world. The air in a tube, such as found in a musical instrument like a flute, can be forced into resonance and produce a pleasant sound, as we discuss in Sound.

At other times, resonance can cause serious problems. A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may vibrate for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often, buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. The span of the roof is also important. Often it is seen that gymnasiums, supermarkets, and churches suffer damage when individual homes suffer far less damage. The roofs with large surface areas supported only at the edges resonate at the frequencies of the earthquakes, causing them to collapse. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged, while areas farther away are damaged.

Summary

- A standing wave is the superposition of two waves which produces a wave that varies in amplitude but does not propagate.
- Nodes are points of no motion in standing waves.
- An antinode is the location of maximum amplitude of a standing wave.
- Normal modes of a wave on a string are the possible standing wave patterns. The lowest frequency that will produce a standing wave is known as the fundamental frequency. The higher frequencies which produce standing waves are called overtones.

Key Equations

Wave speed	$v = \frac{\lambda}{T} = \lambda f$

Linear mass density	$\mu = \frac{\text{mass of the string}}{\text{length of the string}}$
Speed of a wave or pulse on a string under tension	$\lvert v \rvert = \sqrt{\frac{F_T}{\mu}}$
Speed of a compression wave in a fluid	$v = \sqrt{\frac{B}{\rho}}$
Resultant wave from superposition of two sinusoidal waves that are identical except for a phase shift	$y_R\left(x,t\right) = \left[2A\cos\left(\frac{\phi}{2}\right)\right]\sin\left(kx - \omega t + \frac{\phi}{2}\right)$
Wave number	$k \equiv \frac{2\pi}{\lambda}$
Wave speed	$v = \frac{\omega}{k}$
A periodic wave	$y\left(x,t\right) = A\sin\left(kx \mp \omega t + \phi\right)$
Phase of a wave	$kx \mp \omega t + \phi$
The linear wave equation	$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v_w^2}\frac{\partial^2 y(x,t)}{\partial t^2}$
Power averaged over a wavelength	$P_{\text{ave}} = \frac{E_\lambda}{T} = \frac{1}{2}\mu A^2\omega^2\frac{\lambda}{T} = \frac{1}{2}\mu A^2\omega^2 v$
Intensity	$I = \frac{P}{A}$
Intensity for a spherical wave	$I = \frac{P}{4\pi r^2}$
Equation of a standing wave	$y\left(x,t\right) = \left[2A\sin\left(kx\right)\right]\cos\left(\omega t\right)$
Wavelength for symmetric boundary conditions	$\lambda_n = \frac{2}{n}L, \quad n = 1,2,3,4,5...$
Frequency for symmetric boundary conditions	$f_n = n\frac{v}{2L} = nf_1, \quad n = 1,2,3,4,5...$

Conceptual Questions

Exercise:

Problem:

A truck manufacturer finds that a strut in the engine is failing prematurely. A sound engineer determines that the strut resonates at the frequency of the engine and suspects that this could be the problem. What are two possible characteristics of the strut can be modified to correct the problem?

Solution:

It may be as easy as changing the length and/or the density a small amount so that the parts do not resonate at the frequency of the motor.

Exercise:

Problem:

Why do roofs of gymnasiums and churches seem to fail more than family homes when an earthquake occurs?

Exercise:

Problem:

Wine glasses can be set into resonance by moistening your finger and rubbing it around the rim of the glass. Why?

Solution:

Energy is supplied to the glass by the work done by the force of your finger on the glass. When supplied at the right frequency, standing waves form. The glass resonates and the vibrations produce sound.

Exercise:

Problem:

Air conditioning units are sometimes placed on the roof of homes in the city. Occasionally, the air conditioners cause an undesirable hum throughout the upper floors of the homes. Why does this happen? What can be done to reduce the hum?

Exercise:

Problem:

Consider a standing wave modeled as $y\left(x,t\right)=4.00\ \text{cm}\ \text{sin}\left(3\ \text{m}^{-1}x\right)\text{cos}\left(4\ \text{s}^{-1}t\right)$. Is there a node or an antinode at $x=0.00\ \text{m}$? What about a standing wave modeled as $y\left(x,t\right)=4.00\ \text{cm}\ \text{sin}\left(3\ \text{m}^{-1}x+\frac{\pi}{2}\right)\text{cos}\left(4\ \text{s}^{-1}t\right)$? Is there a node or an antinode at the $x=0.00\ \text{m}$ position?

Solution:

For the equation $y\left(x,t\right)=4.00\ \text{cm}\ \text{sin}\left(3\ \text{m}^{-1}x\right)\text{cos}\left(4\ \text{s}^{-1}t\right)$, there is a node because when $x=0.00\ \text{m}$, $\text{sin}\left(3\ \text{m}^{-1}\left(0.00\ \text{m}\right)\right)=0.00$, so $y\left(0.00\ \text{m},t\right)=0.00\ \text{m}$ for all time. For the equation $y\left(x,t\right)=4.00\ \text{cm}\ \text{sin}\left(3\ \text{m}^{-1}x+\frac{\pi}{2}\right)\text{cos}\left(4\ \text{s}^{-1}t\right)$, there is an antinode because when $x=0.00\ \text{m}$, $\text{sin}\left(3\ \text{m}^{-1}\left(0.00\ \text{m}\right)+\frac{\pi}{2}\right)=+1.00$, so $y\left(0.00\ \text{m},t\right)$ oscillates between $+A$ and $-A$ as the cosine term oscillates between +1 and -1.

Problems

Exercise:

Problem:

A wave traveling on a Slinky® that is stretched to 4 m takes 2.4 s to travel the length of the Slinky and back again. (a) What is the speed of the wave? (b) Using the same Slinky stretched to the same length, a standing wave is created which consists of three antinodes and four nodes. At what frequency must the Slinky be oscillating?

Exercise:

Problem:

A 2-m long string is stretched between two supports with a tension that produces a wave speed equal to $v_w = 50.00\,\mathrm{m/s}$. What are the wavelength and frequency of the first three modes that resonate on the string?

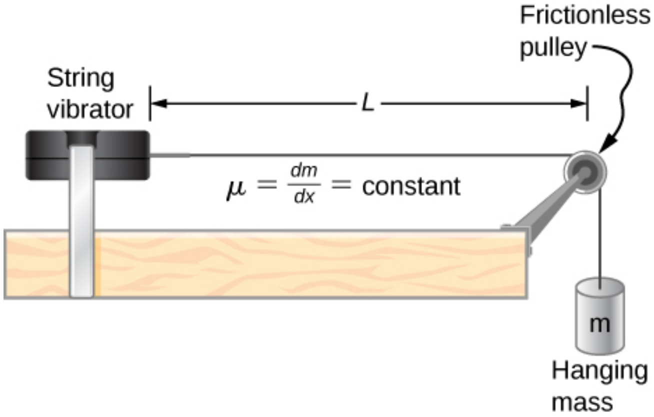
Solution:

$$\lambda_n = \frac{2.00}{n}L, \quad f_n = \frac{v}{\lambda_n}$$
$$\lambda_1 = 4.00\,\mathrm{m}, \quad f_1 = 12.5\,\mathrm{Hz}$$
$$\lambda_2 = 2.00\,\mathrm{m}, \quad f_2 = 25.00\,\mathrm{Hz}$$
$$\lambda_3 = 1.33\,\mathrm{m}, \quad f_3 = 37.59\,\mathrm{Hz}$$

Exercise:

Problem:

Consider the experimental setup shown below. The length of the string between the string vibrator and the pulley is $L = 1.00\,\mathrm{m}$. The linear density of the string is $\mu = 0.006\,\mathrm{kg/m}$. The string vibrator can oscillate at any frequency. The hanging mass is 2.00 kg. (a)What are the wavelength and frequency of $n = 6$ mode? (b) The string oscillates the air around the string. What is the wavelength of the sound if the speed of the sound is $v_s = 343.00\,\mathrm{m/s}$?



Exercise:

Problem:

A cable with a linear density of $\mu = 0.2\,\mathrm{kg/m}$ is hung from telephone poles. The tension in the cable is 500.00 N. The distance between poles is 20 meters. The wind blows across the line, causing the cable resonate. A standing waves pattern is produced that has 4.5 wavelengths between the two poles. The air temperature is $T = 20^\circ\mathrm{C}$. What are the frequency and wavelength of the hum?

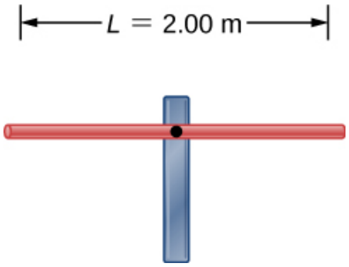
Solution:

$$v = 158.11\,\mathrm{m/s}, \quad \lambda = 4.44\,\mathrm{m}, \quad f = 35.61\,\mathrm{Hz}$$
$$\lambda_s = 9.63\,\mathrm{m}$$

Exercise:

Problem:

Consider a rod of length L, mounted in the center to a support. A node must exist where the rod is mounted on a support, as shown below. Draw the first two normal modes of the rod as it is driven into resonance. Label the wavelength and the frequency required to drive the rod into resonance.



Exercise:

Problem:

Consider two wave functions $y\left(x,t\right)=0.30\,\text{cm}\,\text{sin}\left(3\,\text{m}^{-1}x-4\,\text{s}^{-1}t\right)$ and $y\left(x,t\right)=0.30\,\text{cm}\,\text{sin}\left(3\,\text{m}^{-1}x+4\,\text{s}^{-1}t\right)$. Write a wave function for the resulting standing wave.

Solution:

$y\left(x,t\right)=\left[0.60\,\text{cm}\,\text{sin}\left(3\,\text{m}^{-1}x\right)\right]\text{cos}\left(4\,\text{s}^{-1}t\right)$

Exercise:

Problem:

A 2.40-m wire has a mass of 7.50 g and is under a tension of 160 N. The wire is held rigidly at both ends and set into oscillation. (a) What is the speed of waves on the wire? The string is driven into resonance by a frequency that produces a standing wave with a wavelength equal to 1.20 m. (b) What is the frequency used to drive the string into resonance?

Exercise:

Problem:

A string with a linear mass density of 0.0062 kg/m and a length of 3.00 m is set into the $n=100$ mode of resonance. The tension in the string is 20.00 N. What is the wavelength and frequency of the wave?

Solution:

$\lambda_{100}=0.06\,\text{m}$

$v=56.8\,\text{m/s},\quad f_n=nf_1,\quad n=1,2,3,4,5...$

$f_{100}=947\,\text{Hz}$

Exercise:

Problem:

A string with a linear mass density of 0.0075 kg/m and a length of 6.00 m is set into the $n=4$ mode of resonance by driving with a frequency of 100.00 Hz. What is the tension in the string?

Exercise:

Problem:

Two sinusoidal waves with identical wavelengths and amplitudes travel in opposite directions along a string producing a standing wave. The linear mass density of the string is $\mu = 0.075\,\text{kg/m}$ and the tension in the string is $F_T = 5.00\,\text{N}$. The time interval between instances of total destructive interference is $\Delta t = 0.13\,\text{s}$. What is the wavelength of the waves?

Solution:

$T = 2\Delta t, \quad v = \frac{\lambda}{T}, \quad \lambda = 2.12\,\text{m}$

Exercise:

Problem:

A string, fixed on both ends, is 5.00 m long and has a mass of 0.15 kg. The tension if the string is 90 N. The string is vibrating to produce a standing wave at the fundamental frequency of the string. (a) What is the speed of the waves on the string? (b) What is the wavelength of the standing wave produced? (c) What is the period of the standing wave?

Exercise:

Problem:

A string is fixed at both end. The mass of the string is 0.0090 kg and the length is 3.00 m. The string is under a tension of 200.00 N. The string is driven by a variable frequency source to produce standing waves on the string. Find the wavelengths and frequency of the first four modes of standing waves.

Solution:

$$\begin{aligned}&\lambda_1 = 6.00\,\text{m}, \quad \lambda_2 = 3.00\,\text{m}, \quad \lambda_3 = 2.00\,\text{m}, \quad \lambda_4 = 1.50\,\text{m}\\&v = 258.20\,\text{m/s} = \lambda f\\&f_1 = 43.03\,\text{Hz}, \quad f_2 = 86.07\,\text{Hz}, \quad f_3 = 129.10\,\text{Hz}, \quad f_4 = 172.13\,\text{Hz}\end{aligned}$$

Exercise:

Problem:

The frequencies of two successive modes of standing waves on a string are 258.36 Hz and 301.42 Hz. What is the next frequency above 100.00 Hz that would produce a standing wave?

Exercise:

Problem:

A string is fixed at both ends to supports 3.50 m apart and has a linear mass density of $\mu = 0.005\,\text{kg/m}$. The string is under a tension of 90.00 N. A standing wave is produced on the string with six nodes and five antinodes. What are the wave speed, wavelength, frequency, and period of the standing wave?

Solution:

$v = 134.16\,\text{ms}, \lambda = 1.4\,\text{m}, f = 95.83\,\text{Hz}, T = 0.0104\,\text{s}$

Exercise:

Problem:

Sine waves are sent down a 1.5-m-long string fixed at both ends. The waves reflect back in the opposite direction. The amplitude of the wave is 4.00 cm. The propagation velocity of the waves is 175 m/s. The $n = 6$ resonance mode of the string is produced. Write an equation for the resulting standing wave.

Additional Problems

Exercise:

Problem:

Ultrasound equipment used in the medical profession uses sound waves of a frequency above the range of human hearing. If the frequency of the sound produced by the ultrasound machine is $f = 30\,\text{kHz}$, what is the wavelength of the ultrasound in bone, if the speed of sound in bone is $v = 3000\,\text{m/s}$?

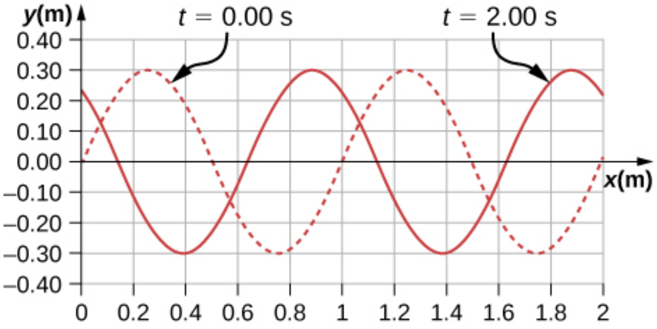
Solution:

$\lambda = 0.10\,\text{m}$

Exercise:

Problem:

Shown below is the plot of a wave function that models a wave at time $t = 0.00\,\text{s}$ and $t = 2.00\,\text{s}$. The dotted line is the wave function at time $t = 0.00\,\text{s}$ and the solid line is the function at time $t = 2.00\,\text{s}$. Estimate the amplitude, wavelength, velocity, and period of the wave.



Exercise:

Problem:

The speed of light in air is approximately $v = 3.00 \times 10^8\,\text{m/s}$ and the speed of light in glass is $v = 2.00 \times 10^8\,\text{m/s}$. A red laser with a wavelength of $\lambda = 633.00\,\text{nm}$ shines light incident of the glass, and some of the red light is transmitted to the glass. The frequency of the light is the same for the air and the glass. (a) What is the frequency of the light? (b) What is the wavelength of the light in the glass?

Solution:

a. $f = 4.74 \times 10^{14}\,\text{Hz}$; b. $\lambda = 422\,\text{nm}$

Exercise:

Problem:

A radio station broadcasts radio waves at a frequency of 101.7 MHz. The radio waves move through the air at approximately the speed of light in a vacuum. What is the wavelength of the radio waves?

Exercise:

Problem:

A sunbather stands waist deep in the ocean and observes that six crests of periodic surface waves pass each minute. The crests are 16.00 meters apart. What is the wavelength, frequency, period, and speed of the waves?

Solution:

$\lambda = 16.00\,\text{m}, \quad f = 0.10\,\text{Hz}, \quad T = 10.00\,\text{s}, \quad v = 1.6\,\text{m/s}$

Exercise:

Problem:

A tuning fork vibrates producing sound at a frequency of 512 Hz. The speed of sound of sound in air is $v = 343.00\,\text{m/s}$ if the air is at a temperature of $20.00\,^\circ\text{C}$. What is the wavelength of the sound?

Exercise:

Problem:

A motorboat is traveling across a lake at a speed of $v_b = 15.00\,\text{m/s}$. The boat bounces up and down every 0.50 s as it travels in the same direction as a wave. It bounces up and down every 0.30 s as it travels in a direction opposite the direction of the waves. What is the speed and wavelength of the wave?

Solution:

$\lambda = (v_b + v)t_b, \quad v = 3.75\,\text{m/s}, \quad \lambda = 3.00\,\text{m}$

Exercise:

Problem:

Use the linear wave equation to show that the wave speed of a wave modeled with the wave function $y\left(x,t\right) = 0.20\,\text{m}\sin\left(3.00\,\text{m}^{-1}x + 6.00\,\text{s}^{-1}t\right)$ is $v = 2.00\,\text{m/s}$. What are the wavelength and the speed of the wave?

Exercise:

Problem:

Given the wave functions $y_1\left(x,t\right) = A\sin\left(kx - \omega t\right)$ and $y_2\left(x,t\right) = A\sin\left(kx - \omega t + \phi\right)$ with $\phi \neq \frac{\pi}{2}$, show that $y_1\left(x,t\right) + y_2\left(x,t\right)$ is a solution to the linear wave equation with a wave velocity of $v = \frac{\omega}{k}$.

Solution:

$$\frac{\partial^2(y_1+y_2)}{\partial t^2} = -A\omega^2 \sin(kx-\omega t) - A\omega^2 \sin(kx-\omega t+\phi)$$
$$\frac{\partial^2(y_1+y_2)}{\partial x^2} = -Ak^2 \sin(kx-\omega t) - Ak^2 \sin(kx-\omega t+\phi)$$
$$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y(x,t)}{\partial t^2}$$
$$-A\omega^2 \sin(kx-\omega t) - A\omega^2 \sin(kx-\omega t+\phi) = \left(\frac{1}{v^2}\right)\left(-Ak^2 \sin(kx-\omega t) - Ak^2 \sin(kx-\omega t+\phi)\right)$$
$$v = \frac{\omega}{k}$$

Exercise:

Problem:

A transverse wave on a string is modeled with the wave function $y(x,t) = 0.10\,\mathrm{m}\sin\left(0.15\,\mathrm{m}^{-1}x + 1.50\,\mathrm{s}^{-1}t + 0.20\right)$. (a) Find the wave velocity. (b) Find the position in the *y*-direction, the velocity perpendicular to the motion of the wave, and the acceleration perpendicular to the motion of the wave, of a small segment of the string centered at $x = 0.40\,\mathrm{m}$ at time $t = 5.00\,\mathrm{s}$.

Exercise:

Problem:

A sinusoidal wave travels down a taut, horizontal string with a linear mass density of $\mu = 0.060\,\mathrm{kg/m}$. The magnitude of maximum vertical acceleration of the wave is $a_{y\,\mathrm{max}} = 0.90\,\mathrm{cm/s}^2$ and the amplitude of the wave is 0.40 m. The string is under a tension of $F_T = 600.00\,\mathrm{N}$. The wave moves in the negative *x*-direction. Write an equation to model the wave.

Solution:

$$y(x,t) = 0.40\,\mathrm{m}\sin\left(0.015\,\mathrm{m}^{-1}x + 1.5\,\mathrm{s}^{-1}t\right)$$

Exercise:

Problem:

A transverse wave on a string $(\mu = 0.0030\,\mathrm{kg/m})$ is described with the equation $y(x,t) = 0.30\,\mathrm{m}\sin\left(\frac{2\pi}{4.00\,\mathrm{m}}\left(x - 16.00\frac{\mathrm{m}}{\mathrm{s}}t\right)\right)$. What is the tension under which the string is held taut?

Exercise:

Problem:

A transverse wave on a horizontal string $(\mu = 0.0060\,\mathrm{kg/m})$ is described with the equation $y(x,t) = 0.30\,\mathrm{m}\sin\left(\frac{2\pi}{4.00\,\mathrm{m}}(x - v_w t)\right)$. The string is under a tension of 300.00 N. What are the wave speed, wave number, and angular frequency of the wave?

Solution:

$$v = 223.61\,\mathrm{m/s},\ k = 1.57\,\mathrm{m}^{-1},\ \omega = 142.43\,\mathrm{s}^{-1}$$

Exercise:

Problem:

A student holds an inexpensive sonic range finder and uses the range finder to find the distance to the wall. The sonic range finder emits a sound wave. The sound wave reflects off the wall and returns to the range finder. The round trip takes 0.012 s. The range finder was calibrated for use at room temperature $T = 20^\circ\text{C}$, but the temperature in the room is actually $T = 23^\circ\text{C}$. Assuming that the timing mechanism is perfect, what percentage of error can the student expect due to the calibration?

Exercise:

Problem:

A wave on a string is driven by a string vibrator, which oscillates at a frequency of 100.00 Hz and an amplitude of 1.00 cm. The string vibrator operates at a voltage of 12.00 V and a current of 0.20 A. The power consumed by the string vibrator is $P = IV$. Assume that the string vibrator is 90% efficient at converting electrical energy into the energy associated with the vibrations of the string. The string is 3.00 m long, and is under a tension of 60.00 N. What is the linear mass density of the string?

Solution:

$$P = \frac{1}{2}A^2(2\pi f)^2\sqrt{\mu F_T}$$
$$\mu = 2.00\ \times\ 10^{-4}\,\text{kg/m}$$

Exercise:

Problem:

A traveling wave on a string is modeled by the wave equation $y\,(x, t) = 3.00\,\text{cm}\,\sin\left(8.00\,\text{m}^{-1}x + 100.00\,\text{s}^{-1}t\right)$. The string is under a tension of 50.00 N and has a linear mass density of $\mu = 0.008\,\text{kg/m}$. What is the average power transferred by the wave on the string?

Exercise:

Problem:

A transverse wave on a string has a wavelength of 5.0 m, a period of 0.02 s, and an amplitude of 1.5 cm. The average power transferred by the wave is 5.00 W. What is the tension in the string?

Solution:

$$P = \frac{1}{2}\mu A^2\omega^2\frac{\lambda}{T},\ \mu = 0.0018\,\text{kg/m}$$

Exercise:

Problem:

(a) What is the intensity of a laser beam used to burn away cancerous tissue that, when 90.0% absorbed, puts 500.0 J of energy into a circular spot 2.00 mm in diameter in 4.00 s? (b) Discuss how this intensity compares to the average intensity of sunlight (about 1kW/m^2) and the implications that would have if the laser beam entered your eye. Note how your answer depends on the time duration of the exposure.

Exercise:

Problem:

Consider two periodic wave functions, $y_1\left(x,t\right)=A\sin\left(kx-\omega t\right)$ and $y_2\left(x,t\right)=A\sin\left(kx-\omega t+\phi\right)$. (a) For what values of ϕ will the wave that results from a superposition of the wave functions have an amplitude of 2*A*? (b) For what values of ϕ will the wave that results from a superposition of the wave functions have an amplitude of zero?

Solution:

a. $A_R=2A\cos\left(\frac{\phi}{2}\right),\ \cos\left(\frac{\phi}{2}\right)=1,\ \phi=0,2\pi,4\pi,$...; b. $A_R=2A\cos\left(\frac{\phi}{2}\right),\ \cos\left(\frac{\phi}{2}\right)=0,\ \phi=0,\pi,3\pi,5\pi$...

Exercise:

Problem:

Consider two periodic wave functions, $y_1\left(x,t\right)=A\sin\left(kx-\omega t\right)$ and $y_2\left(x,t\right)=A\cos\left(kx-\omega t+\phi\right)$. (a) For what values of ϕ will the wave that results from a superposition of the wave functions have an amplitude of 2*A*? (b) For what values of ϕ will the wave that results from a superposition of the wave functions have an amplitude of zero?

Exercise:

Problem:

A trough with dimensions 10.00 meters by 0.10 meters by 0.10 meters is partially filled with water. Small-amplitude surface water waves are produced from both ends of the trough by paddles oscillating in simple harmonic motion. The height of the water waves are modeled with two sinusoidal wave equations, $y_1\left(x,t\right)=0.3\,\text{m}\sin\left(4\,\text{m}^{-1}x-3\,\text{s}^{-1}t\right)$ and $y_2\left(x,t\right)=0.3\,\text{m}\cos\left(4\,\text{m}^{-1}x+3\,\text{s}^{-1}t-\frac{\pi}{2}\right)$. What is the wave function of the resulting wave after the waves reach one another and before they reach the end of the trough (i.e., assume that there are only two waves in the trough and ignore reflections)? Use a spreadsheet to check your results. (*Hint:* Use the trig identities $\sin\left(u\pm v\right)=\sin u\cos v\pm\cos u\sin v$ and $\cos\left(u\pm v\right)=\cos u\cos v\mp\sin u\sin v$)

Solution:

$y_R\left(x,t\right)=0.6\,\text{m}\sin\left(4\,\text{m}^{-1}x\right)\cos\left(3\,\text{s}^{-1}t\right)$

Exercise:

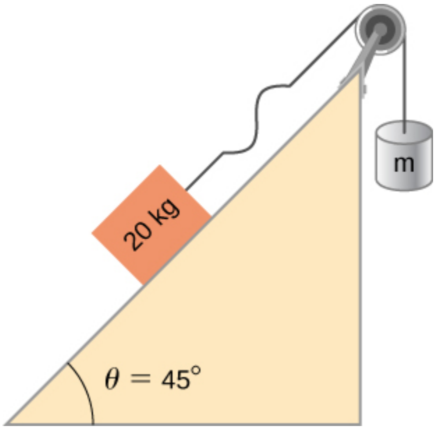
Problem:

A seismograph records the S- and P-waves from an earthquake 20.00 s apart. If they traveled the same path at constant wave speeds of $v_S=4.00\,\text{km/s}$ and $v_P=7.50\,\text{km/s}$, how far away is the epicenter of the earthquake?

Exercise:

Problem:

Consider what is shown below. A 20.00-kg mass rests on a frictionless ramp inclined at 45°. A string with a linear mass density of $\mu=0.025\,\text{kg/m}$ is attached to the 20.00-kg mass. The string passes over a frictionless pulley of negligible mass and is attached to a hanging mass (*m*). The system is in static equilibrium. A wave is induced on the string and travels up the ramp. (a) What is the mass of the hanging mass (*m*)? (b) At what wave speed does the wave travel up the string?



Solution:

a. $$\begin{array}{l}(1) F_T - 20.00\,\text{kg}\left(9.80\,\text{m/s}^2\right)\cos 45^\circ = 0 \\ (2) m\left(9.80\,\text{m/s}^2\right) - F_T = 0 \\ m = 14.14\,\text{kg}\end{array}$$; b. $$\begin{array}{rcl}F_T & = & 138.57\,\text{N} \\ v & = & 67.96\,\text{m/s}\end{array}$$

Exercise:

Problem:

Consider the superposition of three wave functions $y(x,t) = 3.00\,\text{cm}\sin\left(2\,\text{m}^{-1}x - 3\,\text{s}^{-1}t\right)$, $y(x,t) = 3.00\,\text{cm}\sin\left(6\,\text{m}^{-1}x + 3\,\text{s}^{-1}t\right)$, and $y(x,t) = 3.00\,\text{cm}\sin\left(2\,\text{m}^{-1}x - 4\,\text{s}^{-1}t\right)$. What is the height of the resulting wave at position $x = 3.00\,\text{m}$ at time $t = 10.0\,\text{s}$?

Exercise:

Problem:

A string has a mass of 150 g and a length of 3.4 m. One end of the string is fixed to a lab stand and the other is attached to a spring with a spring constant of $k_s = 100\,\text{N/m}$. The free end of the spring is attached to another lab pole. The tension in the string is maintained by the spring. The lab poles are separated by a distance that stretches the spring 2.00 cm. The string is plucked and a pulse travels along the string. What is the propagation speed of the pulse?

Solution:

$F_T = 12\,\text{N}, v = 16.49\,\text{m/s}$

Exercise:

Problem:

A standing wave is produced on a string under a tension of 70.0 N by two sinusoidal transverse waves that are identical, but moving in opposite directions. The string is fixed at $x = 0.00\,\text{m}$ and $x = 10.00\,\text{m}$. Nodes appear at $x = 0.00\,\text{m}$, 2.00 m, 4.00 m, 6.00 m, 8.00 m, and 10.00 m. The amplitude of the standing wave is 3.00 cm. It takes 0.10 s for the antinodes to make one complete oscillation. (a) What are the wave functions of the two sine waves that produce the standing wave? (b) What are the maximum velocity and acceleration of the string, perpendicular to the direction of motion of the transverse waves, at the antinodes?

Exercise:

Problem:

A string with a length of 4 m is held under a constant tension. The string has a linear mass density of $\mu = 0.006\,\text{kg/m}$. Two resonant frequencies of the string are 400 Hz and 480 Hz. There are no resonant frequencies between the two frequencies. (a) What are the wavelengths of the two resonant modes? (b) What is the tension in the string?

Solution:

a. $f_n = \frac{nv}{2L},\ v = \frac{2Lf_{n+1}}{n+1},\ \frac{n+1}{n} = \frac{2Lf_{n+1}}{2Lf_n},\ 1 + \frac{1}{n} = 1.2,\ n = 5$; b. $F_T = 245.76\,\text{N}$
$\lambda_n = \frac{2}{n}L,\ \lambda_5 = 1.6\,\text{m},\ \lambda_6 = 1.33\,\text{m}$

Challenge Problems

Exercise:

Problem:

A copper wire has a radius of 200 μm and a length of 5.0 m. The wire is placed under a tension of 3000 N and the wire stretches by a small amount. The wire is plucked and a pulse travels down the wire. What is the propagation speed of the pulse? (Assume the temperature does not change: $\left(\rho = 8.96\frac{\text{g}}{\text{cm}^3}, Y = 1.1 \times 10^{11}\frac{\text{N}}{\text{m}}\right).$)

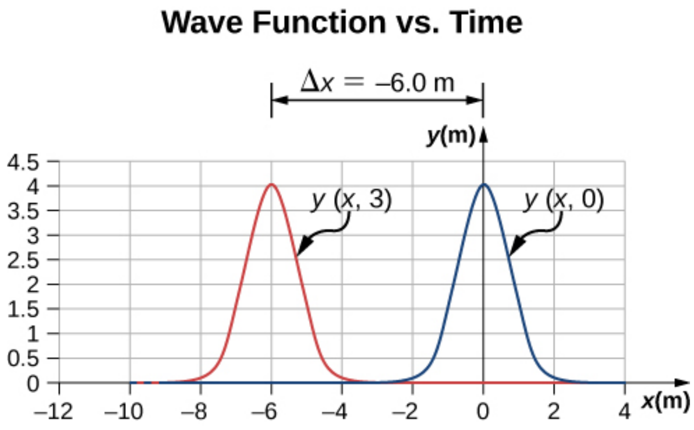
Exercise:

Problem:

A pulse moving along the *x* axis can be modeled as the wave function $y\left(x,t\right) = 4.00\,\text{m}e^{-\left(\frac{x+(2.00\,\text{m/s})t}{1.00\,\text{m}}\right)^2}$. (a)What are the direction and propagation speed of the pulse? (b) How far has the wave moved in 3.00 s? (c) Plot the pulse using a spreadsheet at time $t = 0.00\,\text{s}$ and $t = 3.00\,\text{s}$ to verify your answer in part (b).

Solution:

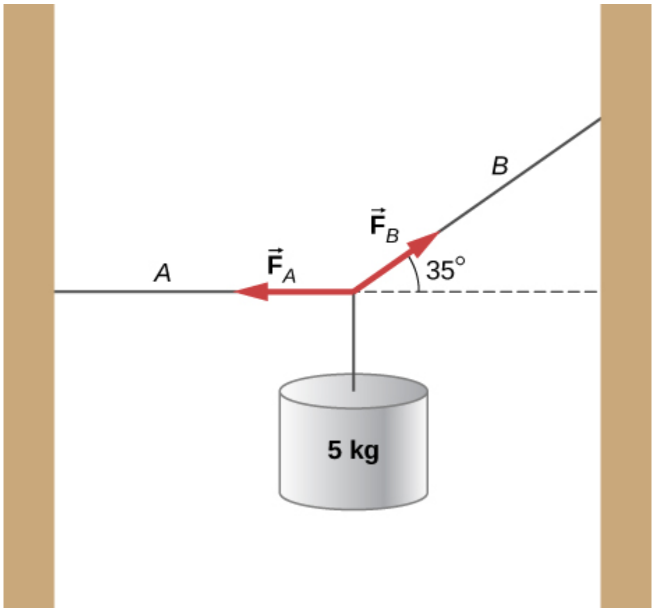
a. Moves in the negative *x* direction at a propagation speed of $v = 2.00\,\text{m/s}$. b. $\Delta x = -6.00\,\text{m}$; c.


Wave Function vs. Time

Exercise:

Problem:

A string with a linear mass density of $\mu = 0.0085\ \mathrm{kg/m}$ is fixed at both ends. A 5.0-kg mass is hung from the string, as shown below. If a pulse is sent along section A, what is the wave speed in section A and the wave speed in section B?



Exercise:

Problem:

Consider two wave functions $y_1(x,t) = A\sin(kx - \omega t)$ and $y_2(x,t) = A\sin(kx + \omega t + \phi)$. What is the wave function resulting from the interference of the two wave? (*Hint:* $\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$ and $\phi = \frac{\phi}{2} + \frac{\phi}{2}$.)

Solution:

$$\sin(kx - \omega t) = \sin\left(kx + \frac{\phi}{2}\right)\cos\left(\omega t + \frac{\phi}{2}\right) - \cos\left(kx + \frac{\phi}{2}\right)\sin\left(\omega t + \frac{\phi}{2}\right)$$
$$\sin(kx - \omega t + \phi) = \sin\left(kx + \frac{\phi}{2}\right)\cos\left(\omega t + \frac{\phi}{2}\right) + \cos\left(kx + \frac{\phi}{2}\right)\sin\left(\omega t + \frac{\phi}{2}\right)$$
$$\sin(kx - \omega t) + \sin(kx + \omega t + \phi) = 2\sin\left(kx + \frac{\phi}{2}\right)\cos\left(\omega t + \frac{\phi}{2}\right)$$
$$y_R = 2\,A\sin\left(kx + \frac{\phi}{2}\right)\cos\left(\omega t + \frac{\phi}{2}\right)$$

Exercise:

Problem:

The wave function that models a standing wave is given as $y_R(x,t) = 6.00\ \mathrm{cm}\sin\left(3.00\ \mathrm{m}^{-1}x + 1.20\ \mathrm{rad}\right)\cos\left(6.00\ \mathrm{s}^{-1}t + 1.20\ \mathrm{rad}\right)$. What are two wave functions that interfere to form this wave function? Plot the two wave functions and the sum of the sum of the two wave functions at $t = 1.00\ \mathrm{s}$ to verify your answer.

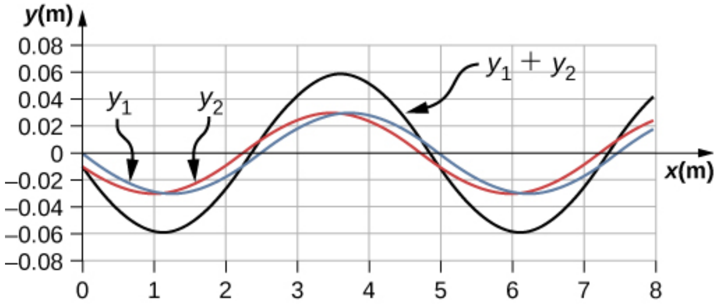
Exercise:

Problem:

Consider two wave functions $y_1(x,t) = A\sin(kx - \omega t)$ and $y_2(x,t) = A\sin(kx + \omega t + \phi)$. The resultant wave form when you add the two functions is $y_R = 2A\sin\left(kx + \frac{\phi}{2}\right)\cos\left(\omega t + \frac{\phi}{2}\right)$. Consider the case where $A = 0.03\,\text{m}^{-1}$, $k = 1.26\,\text{m}^{-1}$, $\omega = \pi\,\text{s}^{-1}$, and $\phi = \frac{\pi}{10}$. (a) Where are the first three nodes of the standing wave function starting at zero and moving in the positive *x* direction? (b) Using a spreadsheet, plot the two wave functions and the resulting function at time $t = 1.00\,\text{s}$ to verify your answer.

Solution:

$\sin\left(kx + \frac{\phi}{2}\right) = 0,\ kx + \frac{\phi}{2} = 0, \pi, 2\pi,\ 1.26\,\text{m}^{-1}x + \frac{\pi}{20} = \pi, 2\pi, 3\pi;$
$x = 2.37\,\text{m}, 4.86\,\text{m}, 7.35\,\text{m}$



Glossary

antinode
location of maximum amplitude in standing waves

fundamental frequency
lowest frequency that will produce a standing wave

node
point where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

normal mode
possible standing wave pattern for a standing wave on a string

overtone
frequency that produces standing waves and is higher than the fundamental frequency

standing wave
wave that can bounce back and forth through a particular region, effectively becoming stationary

Introduction class="introduction"

Hearing is an important human sense that can detect frequencies of sound, ranging between 20 Hz and 20 kHz. However, other species have very different ranges of hearing. Bats, for example, emit clicks in ultrasound, using frequencies beyond 20 kHz. They can detect nearby insects by hearing the echo of these ultrasonic clicks. Ultrasound is important in several human applications, including probing the interior structures of human bodies, Earth, and the Sun. Ultrasound is also useful in industry for nondestructive testing. (credit: modification of work by Angell Williams)

Sound is an example of a mechanical wave, specifically, a pressure wave: Sound waves travel through the air and other media as oscillations of molecules. Normal human hearing encompasses an impressive range of frequencies from 20 Hz to 20 kHz. Sounds below 20 Hz are called infrasound, whereas those above 20 kHz are called ultrasound. Some animals, like the bat shown in [link], can hear sounds in the ultrasonic range.

Many of the concepts covered in Waves also have applications in the study of sound. For example, when a sound wave encounters an interface between two media with different wave speeds, reflection and transmission of the wave occur.

Ultrasound has many uses in science, engineering, and medicine. Ultrasound is used for nondestructive testing in engineering, such as testing the thickness of coating on metal. In medicine, sound waves are far less destructive than X-rays and can be used to image the fetus in a mother’s womb without danger to the fetus or the mother. Later in this chapter, we discuss the Doppler effect, which can be used to determine the velocity of blood in the arteries or wind speed in weather systems.

Sound Waves

By the end of this section, you will be able to:

- Explain the difference between sound and hearing
- Describe sound as a wave
- List the equations used to model sound waves
- Describe compression and rarefactions as they relate to sound

The physical phenomenon of **sound** is a disturbance of matter that is transmitted from its source outward. **Hearing** is the perception of sound, just as seeing is the perception of visible light. On the atomic scale, sound is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. Thus, sound waves can induce oscillations and resonance effects ([link]).

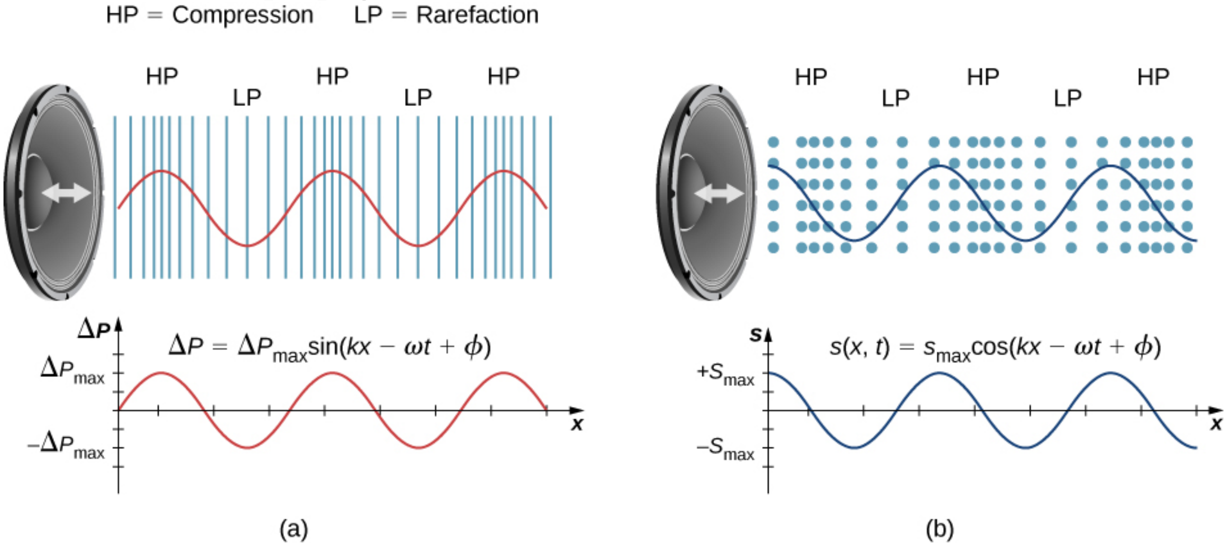
This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. (credit: “||read||”/Flickr)

Note:
This video shows waves on the surface of a wine glass, being driven by sound waves from a speaker. As the frequency of the sound wave approaches the resonant frequency of the wine glass, the amplitude and frequency of the waves on the wine glass increase. When the resonant frequency is reached, the glass shatters.

A speaker produces a sound wave by oscillating a cone, causing vibrations of air molecules. In [link], a speaker vibrates at a constant frequency and amplitude, producing vibrations in the surrounding air molecules. As the speaker oscillates back and forth, it transfers energy to the air, mostly as thermal energy. But a small part of the speaker’s energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high-pressure regions) and rarefactions (low-pressure regions) move out as longitudinal pressure waves having the same frequency as the speaker—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.)

[link](a) shows the compressions and rarefactions, and also shows a graph of gauge pressure versus distance from a speaker. As the speaker moves in the positive *x*-direction, it pushes air molecules, displacing them from their equilibrium positions. As the speaker moves in the negative *x*-direction, the air molecules move back toward their equilibrium positions due to a restoring force. The air molecules oscillate in simple harmonic motion about their equilibrium positions, as shown in part (b). Note that sound waves in air are longitudinal, and in the figure, the wave propagates in the

positive *x*-direction and the molecules oscillate parallel to the direction in which the wave propagates.


HP = Compression
LP = Rarefaction
HP
LP
HP
LP
HP
ΔP
$\Delta P = \Delta P_{max} \sin(kx - \omega t + \phi)$
ΔP_{max}
$-\Delta P_{max}$
x
(a)
HP
LP
HP
LP
HP
s
$s(x, t) = s_{max} \cos(kx - \omega t + \phi)$
$+S_{max}$
$-S_{max}$
x
(b)

(a) A vibrating cone of a speaker, moving in the positive *x*-direction, compresses the air in front of it and expands the air behind it. As the speaker oscillates, it creates another compression and rarefaction as those on the right move away from the speaker. After many vibrations, a series of compressions and rarefactions moves out from the speaker as a sound wave. The red graph shows the gauge pressure of the air versus the distance from the speaker. Pressures vary only slightly from atmospheric pressure for ordinary sounds. Note that gauge pressure is modeled with a sine function, where the crests of the function line up with the compressions and the troughs line up with the rarefactions. (b) Sound waves can also be modeled using the displacement of the air molecules. The blue graph shows the displacement of the air molecules versus the position from the speaker and is modeled with a cosine function. Notice that the displacement is zero for the molecules in their equilibrium position and are centered at the compressions and rarefactions. Compressions are formed when molecules on either side of the equilibrium molecules are displaced toward the equilibrium position. Rarefactions are formed when the molecules are displaced away from the equilibrium position.

Note:
Models Describing Sound
Sound can be modeled as a pressure wave by considering the change in pressure from average pressure,
Equation:

$$\Delta P = \Delta P_{\text{max}} \sin\left(kx \mp \omega t + \phi\right).$$

This equation is similar to the periodic wave equations seen in Waves, where ΔP is the change in pressure, ΔP_{max} is the maximum change in pressure, $k = \frac{2\pi}{\lambda}$ is the wave number, $\omega = \frac{2\pi}{T} = 2\pi f$ is the angular frequency, and ϕ is the initial phase. The wave speed can be determined from $v = \frac{\omega}{k} = \frac{\lambda}{T}$. Sound waves can also be modeled in terms of the displacement of the air molecules. The displacement of the air molecules can be modeled using a cosine function:
Equation:

$$s\left(x, t\right) = s_{\text{max}} \cos\left(kx \mp \omega t + \phi\right).$$

In this equation, s is the displacement and s_{max} is the maximum displacement.

Not shown in the figure is the amplitude of a sound wave as it decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. The intensity decreases as it moves away from the speaker, as discussed in Waves. The energy is also absorbed by objects and converted into thermal energy by the viscosity of the air. In addition, during each compression, a little heat transfers to the air; during each rarefaction, even less heat transfers from the air, and these heat transfers reduce the organized disturbance into random thermal motions. Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important characteristics for sound, as they are for all waves.

Summary

- Sound is a disturbance of matter (a pressure wave) that is transmitted from its source outward. Hearing is the perception of sound.
- Sound can be modeled in terms of pressure or in terms of displacement of molecules.
- The human ear is sensitive to frequencies between 20 Hz and 20 kHz.

Conceptual Questions

Exercise:

Problem: What is the difference between sound and hearing?

Solution:

Sound is a disturbance of matter (a pressure wave) that is transmitted from its source outward. Hearing is the human perception of sound.

Exercise:

Problem:

You will learn that light is an electromagnetic wave that can travel through a vacuum. Can sound waves travel through a vacuum?

Exercise:

Problem:

Sound waves can be modeled as a change in pressure. Why is the change in pressure used and not the actual pressure?

Solution:

Consider a sound wave moving through air. The pressure of the air is the equilibrium condition, it is the change in pressure that produces the sound wave.

Problems

Exercise:

Problem:

Consider a sound wave modeled with the equation $s\left(x,t\right)=4.00\,\text{nm}\cos\left(3.66\,\text{m}^{-1}\,x-1256\,\text{s}^{-1}t\right)$. What is the maximum displacement, the wavelength, the frequency, and the speed of the sound wave?

Solution:

$s_{\text{max}}=4.00\,\text{nm},\quad\lambda=1.72\,\text{m},\quad f=200\,\text{Hz},\quad v=343.17\,\text{m/s}$

Exercise:

Problem:

Consider a sound wave moving through the air modeled with the equations$s\left(x,t\right)=6.00\,\text{nm}\cos\left(54.93\,\text{m}^{-1}x-18.84\;\times\;10^{3}\,\text{s}^{-1}t\right)$. What is the shortest time required for an air molecule to move between 3.00 nm and –3.00 nm?

Exercise:

Problem:

Consider a diagnostic ultrasound of frequency 5.00 MHz that is used to examine an irregularity in soft tissue. (a) What is the wavelength in air of such a sound wave if the speed of sound is 343 m/s? (b) If the speed of sound in tissue is 1800 m/s, what is the wavelength of this wave in tissue?

Solution:

a. $\lambda=68.60\;\mu\text{m}$; b. $\lambda=360.00\;\mu\text{m}$

Exercise:

Problem:

A sound wave is modeled as $\Delta P = 1.80\,\text{Pa}\sin\left(55.41\,\text{m}^{-1}\,x - 18{,}840\,\text{s}^{-1}t\right)$. What is the maximum change in pressure, the wavelength, the frequency, and the speed of the sound wave?

Exercise:

Problem:

A sound wave is modeled with the wave function $\Delta P = 1.20\,\text{Pa}\sin\left(kx - 6.28\,\times\,10^4\text{s}^{-1}t\right)$ and the sound wave travels in air at a speed of $v = 343.00\,\text{m/s}$. (a) What is the wave number of the sound wave? (b) What is the value for $\Delta P\,(3.00\,\text{m}, 20.00\,\text{s})$?

Solution:

a. $k = 183.09\text{m}^{-1}$;
b. $\Delta P = -1.11\,\text{Pa}$

Exercise:

Problem:

The displacement of the air molecules in sound wave is modeled with the wave function $s\,(x, t) = 5.00\,\text{nm}\cos\left(91.54\,\text{m}^{-1}x - 3.14\,\times\,10^4\,\text{s}^{-1}t\right)$. (a) What is the wave speed of the sound wave? (b) What is the maximum speed of the air molecules as they oscillate in simple harmonic motion? (c) What is the magnitude of the maximum acceleration of the air molecules as they oscillate in simple harmonic motion?

Exercise:

Problem:

A speaker is placed at the opening of a long horizontal tube. The speaker oscillates at a frequency *f*, creating a sound wave that moves down the tube. The wave moves through the tube at a speed of $v = 340.00\,\text{m/s}$. The sound wave is modeled with the wave function $s\left(x,t\right) = s_{\text{max}}\text{cos}\left(kx - \omega t + \phi\right)$. At time $t = 0.00\,\text{s}$, an air molecule at $x = 3.5\,\text{m}$ is at the maximum displacement of 7.00 nm. At the same time, another molecule at $x = 3.7\,\text{m}$ has a displacement of 3.00 nm. What is the frequency at which the speaker is oscillating?

Solution:

$$\begin{gathered} s_1 = 7.00\,\text{nm},\;\; s_2 = 3.00\,\text{nm},\;\; kx_1 + \phi = 0\,\text{rad} \\ kx_2 + \phi = 1.128\,\text{rad} \\ k\left(x_2 - x_1\right) = 1.128\,\text{rad},\;\; k = 5.64\,\text{m}^{-1} \\ \lambda = 1.11\,\text{m},\;\; f = 306.31\,\text{Hz} \end{gathered}$$

Exercise:

Problem:

A 250-Hz tuning fork is struck and begins to vibrate. A sound-level meter is located 34.00 m away. It takes the sound $\Delta t = 0.10\,\text{s}$ to reach the meter. The maximum displacement of the tuning fork is 1.00 mm. Write a wave function for the sound.

Exercise:

Problem:

A sound wave produced by an ultrasonic transducer, moving in air, is modeled with the wave equation $s\left(x,t\right) = 4.50\,\text{nm}\,\text{cos}\left(9.15\;\times\;10^4\,\text{m}^{-1}x - 2\pi\left(5.00\,\text{MHz}\right)t\right)$. The transducer is to be used in nondestructive testing to test for fractures in steel beams. The speed of sound in the steel beam is $v = 5950\,\text{m/s}$. Find the wave function for the sound wave in the steel beam.

Solution:

$k=5.28\ \times\ 10^{3}\mathrm{m}$
$s\left(x,t\right)=4.50\,\mathrm{nm}\cos\left(5.28\ \times\ 10^{3}\,\mathrm{m}^{-1}x-2\pi\left(5.00\,\mathrm{MHz}\right)t\right)$

Exercise:

Problem:

Porpoises emit sound waves that they use for navigation. If the wavelength of the sound wave emitted is 4.5 cm, and the speed of sound in the water is $v=1530\,\mathrm{m/s}$, what is the period of the sound?

Exercise:

Problem:

Bats use sound waves to catch insects. Bats can detect frequencies up to 100 kHz. If the sound waves travel through air at a speed of $v=343\,\mathrm{m/s}$, what is the wavelength of the sound waves?

Solution:

$\lambda=3.43\,\mathrm{mm}$

Exercise:

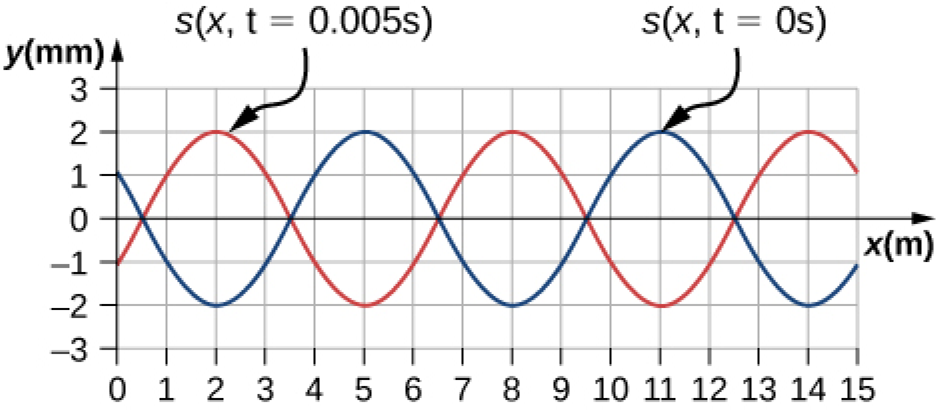
Problem:

A bat sends of a sound wave 100 kHz and the sound waves travel through air at a speed of $v=343\,\mathrm{m/s}$. (a) If the maximum pressure difference is 1.30 Pa, what is a wave function that would model the sound wave, assuming the wave is sinusoidal? (Assume the phase shift is zero.) (b) What are the period and wavelength of the sound wave?

Exercise:

Problem:

Consider the graph shown below of a compression wave. Shown are snapshots of the wave function for $t=0.000\,\text{s}$ (blue) and $t=0.005\,\text{s}$ (orange). What are the wavelength, maximum displacement, velocity, and period of the compression wave?



Solution:

$\lambda=6.00\,\text{m}$

$s_{\text{max}}=2.00\,\text{mm}$

$v=600\,\text{m/s}$

$T=0.01\,\text{s}$

Exercise:

Problem:

Consider the graph in the preceding problem of a compression wave. Shown are snapshots of the wave function for $t=0.000\,\text{s}$ (blue) and $t=0.005\,\text{s}$ (orange). Given that the displacement of the molecule at time $t=0.00\,\text{s}$ and position $x=0.00\,\text{m}$ is $s\left(0.00\,\text{m},0.00\,\text{s}\right)=1.08\,\text{mm}$, derive a wave function to model the compression wave.

Exercise:

Problem:

A guitar string oscillates at a frequency of 100 Hz and produces a sound wave. (a) What do you think the frequency of the sound wave is that the vibrating string produces? (b) If the speed of the sound wave is $v = 343\,\mathrm{m/s}$, what is the wavelength of the sound wave?

Solution:

(a) $f = 100\,\mathrm{Hz}$, (b) $\lambda = 3.43\,\mathrm{m}$

Glossary

hearing
  perception of sound

sound
  traveling pressure wave that may be periodic; the wave can be modeled as a pressure wave or as an oscillation of molecules

Speed of Sound

By the end of this section, you will be able to:

- Explain the relationship between wavelength and frequency of sound
- Determine the speed of sound in different media
- Derive the equation for the speed of sound in air
- Determine the speed of sound in air for a given temperature

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display ([link]). You see the flash of an explosion well before you hear its sound and possibly feel the pressure wave, implying both that sound travels at a finite speed and that it is much slower than light.

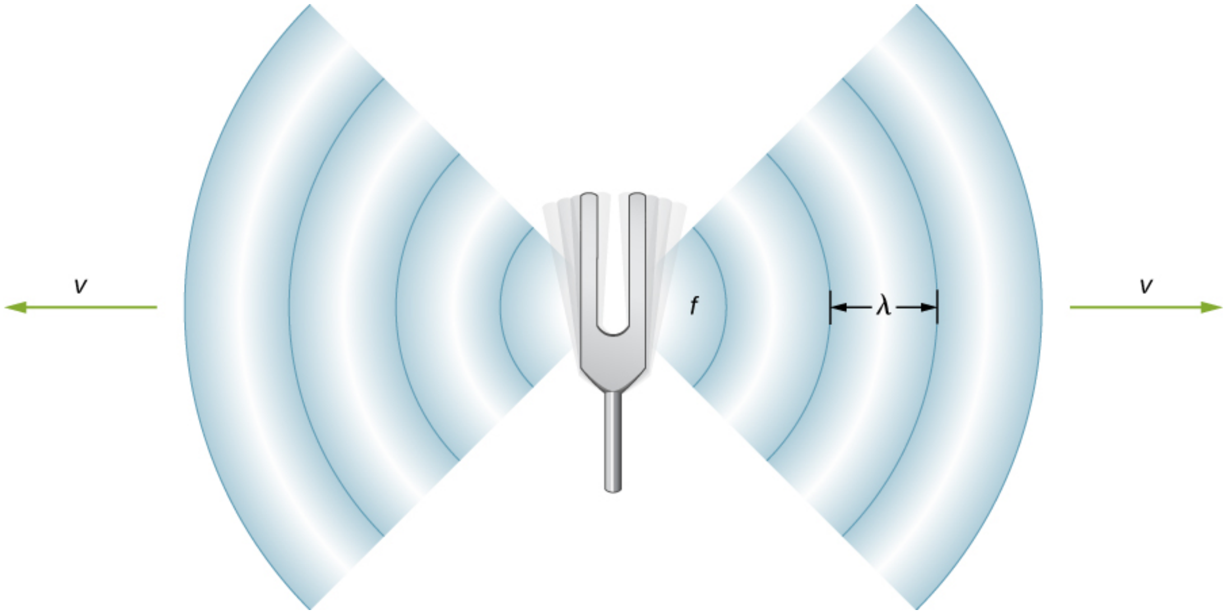
When a firework shell explodes, we perceive the light energy before the sound energy because sound travels more slowly than light does.

The difference between the speed of light and the speed of sound can also be experienced during an electrical storm. The flash of lighting is often seen before the clap of thunder. You may have heard that if you count the number of seconds between the flash and the sound, you can estimate the distance to the source. Every five seconds converts to about one mile. The velocity of any wave is related to its frequency and wavelength by

Note:
Equation:

$$v = f\lambda,$$

where v is the speed of the wave, f is its frequency, and λ is its wavelength. Recall from Waves that the wavelength is the length of the wave as measured between sequential identical points. For example, for a surface water wave or sinusoidal wave on a string, the wavelength can be measured between any two convenient sequential points with the same height and slope, such as between two sequential crests or two sequential troughs. Similarly, the wavelength of a sound wave is the distance between sequential identical parts of a wave—for example, between sequential compressions ([link]). The frequency is the same as that of the source and is the number of waves that pass a point per unit time.



A sound wave emanates from a source, such as a tuning fork, vibrating at a frequency *f*. It propagates at speed *v* and has a wavelength λ.

Speed of Sound in Various Media

[link] shows that the speed of sound varies greatly in different media. The speed of sound in a medium depends on how quickly vibrational energy can be transferred through the medium. For this reason, the derivation of the speed of sound in a medium depends on the medium and on the state of the medium. In general, the equation for the speed of a mechanical wave in a medium depends on the square root of the restoring force, or the elastic property, divided by the inertial property,

Equation:

$$v=\sqrt{\frac{\text{elastic property}}{\text{inertial property}}}.$$

Also, sound waves satisfy the wave equation derived in Waves,

Equation:

$$\frac{\partial^2 y\left(x,t\right)}{\partial x^2}=\frac{1}{v^2}\frac{\partial^2 y\left(x,t\right)}{\partial t^2}.$$

Recall from Waves that the speed of a wave on a string is equal to $v=\sqrt{\frac{F_T}{\mu}}$, where the restoring force is the tension in the string F_T and the linear density μ is the inertial property. In a fluid, the speed of sound depends on the bulk modulus and the density,

Note:
Equation:

$$v=\sqrt{\frac{B}{\rho}}.$$

The speed of sound in a solid the depends on the Young's modulus of the medium and the density,

Note:
Equation:

$$v=\sqrt{\frac{Y}{\rho}}.$$

In an ideal gas (see The Kinetic Theory of Gases), the equation for the speed of sound is

Note:
Equation:

$$v=\sqrt{\frac{\gamma R T_{\mathrm{K}}}{M}},$$

where γ is the adiabatic index, $R=8.31\ \mathrm{J/mol\cdot K}$ is the gas constant, T_{K} is the absolute temperature in kelvins, and M is the molecular mass. In general, the more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of simple harmonic motion is directly proportional to the stiffness of the oscillating object as measured by k, the spring constant. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to m, the mass of the oscillating object. The speed of sound in air is low, because air is easily compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Medium	***v*** **(m/s)**
Gases at $0^{\circ}C$	
Air	331
Carbon dioxide	259
Oxygen	316

Medium	***v* (m/s)**
Helium	965
Hydrogen	1290
Liquids at $20^{\circ}C$	
Ethanol	1160
Mercury	1450
Water, fresh	1480
Sea Water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

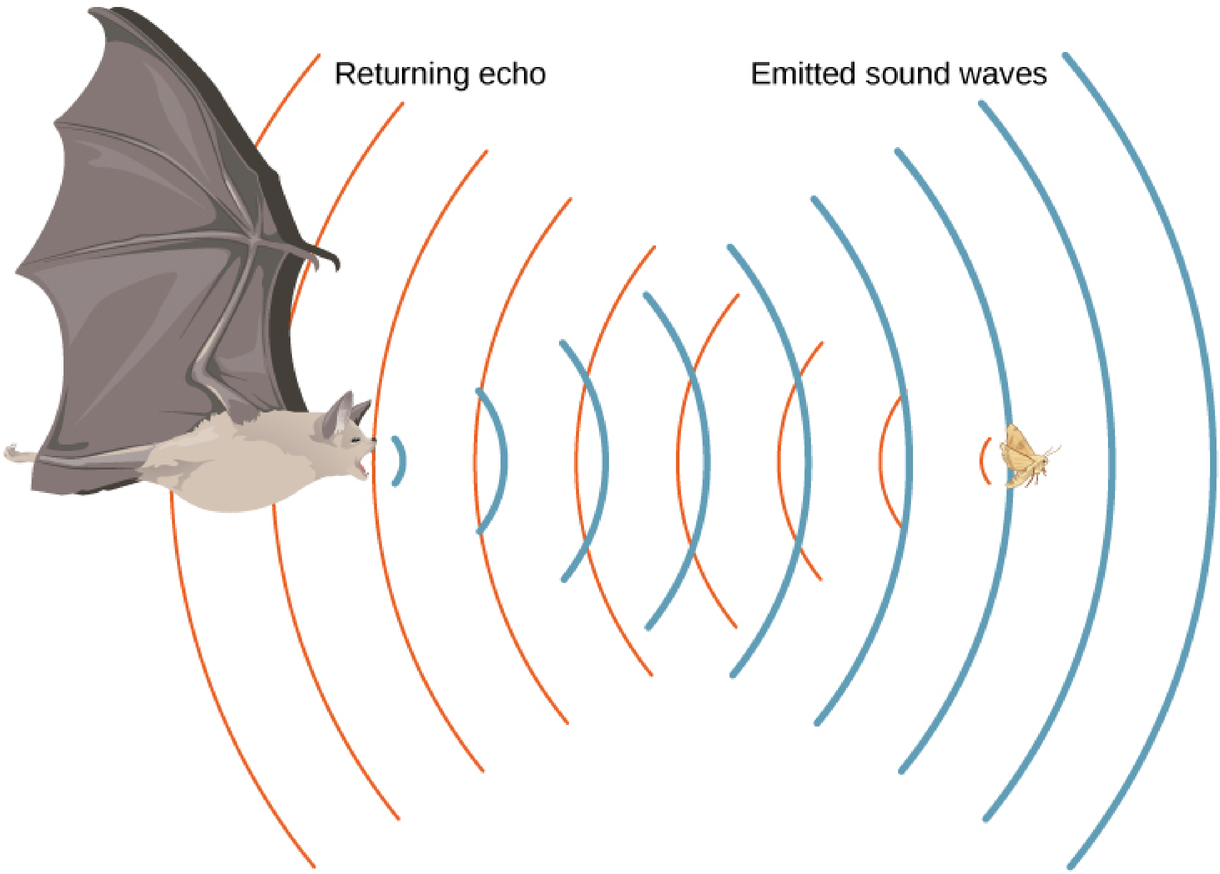
Speed of Sound in Various Media

Because the speed of sound depends on the density of the material, and the density depends on the temperature, there is a relationship between the temperature in a given medium and the speed of sound in the medium. For air at sea level, the speed of sound is given by

Note:
Equation:

$$v = 331\frac{\text{m}}{\text{s}}\sqrt{1+\frac{T_{\text{C}}}{273^\circ\text{C}}} = 331\frac{\text{m}}{\text{s}}\sqrt{\frac{T_{\text{K}}}{273\,\text{K}}}$$

where the temperature in the first equation (denoted as T_{C}) is in degrees Celsius and the temperature in the second equation (denoted as T_{K}) is in kelvins. The speed of sound in gases is related to the average speed of particles in the gas, $v_{\text{rms}} = \sqrt{\frac{3k_{\text{B}}T}{m}}$, where k_{B} is the Boltzmann constant $(1.38 \times 10^{-23}\,\text{J/K})$ and *m* is the mass of each (identical) particle in the gas. Note that *v* refers to the speed of the coherent propagation of a disturbance (the wave), whereas v_{rms} describes the speeds of particles in random directions. Thus, it is reasonable that the speed of sound in air and other gases should depend on the square root of temperature. While not negligible, this is not a strong dependence. At 0°C, the speed of sound is 331 m/s, whereas at 20.0°C, it is 343 m/s, less than a 4% increase. [link] shows how a bat uses the speed of sound to sense distances.



A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

Derivation of the Speed of Sound in Air

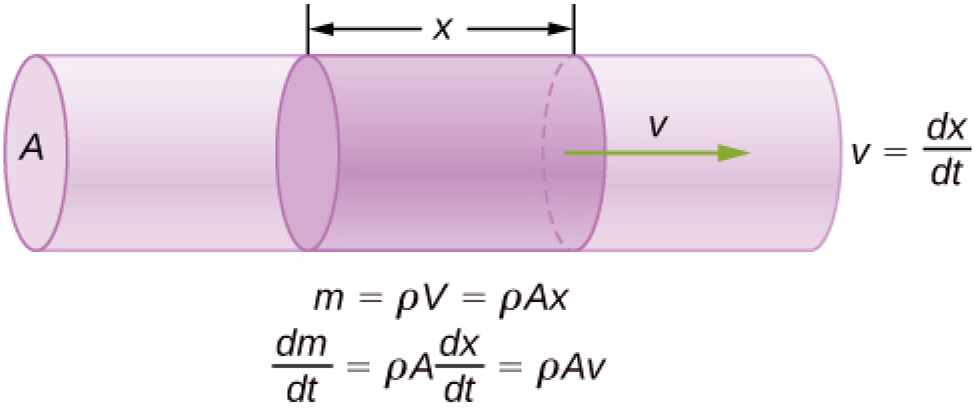
As stated earlier, the speed of sound in a medium depends on the medium and the state of the medium. The derivation of the equation for the speed of sound in air starts with the mass flow rate and continuity equation discussed in Fluid Mechanics.

Consider fluid flow through a pipe with cross-sectional area A ([link]). The mass in a small volume of length x of the pipe is equal to the density times the volume, or $m = \rho V = \rho A x$. The mass flow rate is

Equation:

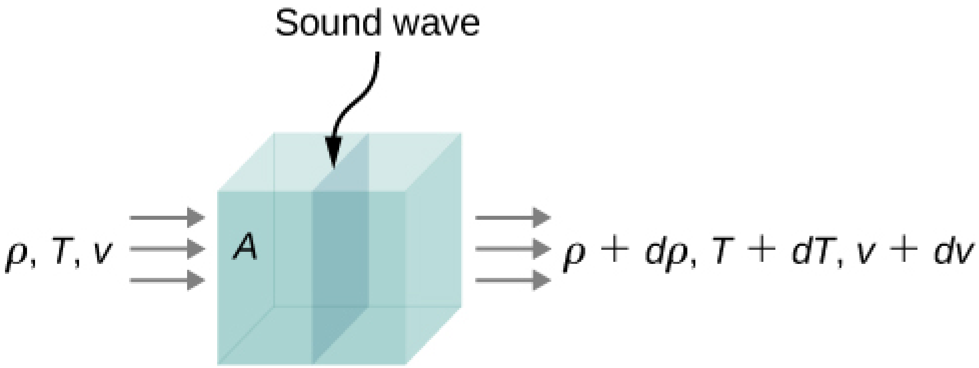
$$\frac{dm}{dt} = \frac{d}{dt}(\rho V) = \frac{d}{dt}(\rho Ax) = \rho A\frac{dx}{dt} = \rho Av.$$

The continuity equation from Fluid Mechanics states that the mass flow rate into a volume has to equal the mass flow rate out of the volume, $\rho_{\text{in}}A_{\text{in}}v_{\text{in}} = \rho_{\text{out}}A_{\text{out}}v_{\text{out}}$.



The mass of a fluid in a volume is equal to the density times the volume, $m = \rho V = \rho Ax$. The mass flow rate is the time derivative of the mass.

Now consider a sound wave moving through a parcel of air. A parcel of air is a small volume of air with imaginary boundaries ([link]). The density, temperature, and velocity on one side of the volume of the fluid are given as ρ, T, v, and on the other side are $\rho + d\rho, T + dT, v + dv$.



A sound wave moves through a volume of fluid. The density, temperature, and velocity of the fluid change from one side to the other.

The continuity equation states that the mass flow rate entering the volume is equal to the mass flow rate leaving the volume, so

Equation:

$$\rho A v = (\rho + d\rho) A \,(v + dv).$$

This equation can be simplified, noting that the area cancels and considering that the multiplication of two infinitesimals is approximately equal to zero: $d\rho\,(dv) \approx 0$,

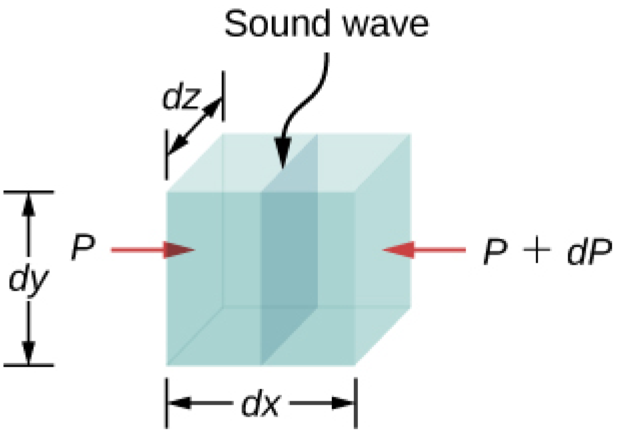
Equation:

$$\begin{aligned} \rho v &= (\rho + d\rho)\,(v + dv) \\ \rho v &= \rho v + \rho\,(dv) + (d\rho) v + (d\rho)\,(dv) \\ 0 &= \rho\,(dv) + (d\rho) v \\ \rho\, dv &= -v d\rho. \end{aligned}$$

The net force on the volume of fluid ([link]) equals the sum of the forces on the left face and the right face:

Equation:

$$\begin{aligned}F_{\text{net}} &= p\,dy\,dz - (p + dp)dy\,dz \\ &= p\,dy\,dz - pdy\,dz - dp\,dy\,dz \\ &= -dp\,dy\,dz \\ ma &= -dp\,dy\,dz.\end{aligned}$$



A sound wave moves through a volume of fluid. The force on each face can be found by the pressure times the area.

The acceleration is the force divided by the mass and the mass is equal to the density times the volume, $m = \rho V = \rho\, dx\, dy\, dz$. We have

Equation:

$$\begin{aligned}ma &= -dp\,dy\,dz \\ a &= -\frac{dp\,dy\,dz}{m} = -\frac{dp\,dy\,dz}{\rho\,dx\,dy\,dz} = -\frac{dp}{(\rho\,dx)} \\ \frac{dv}{dt} &= -\frac{dp}{(\rho\,dx)} \\ dv &= -\frac{dp}{(\rho\,dx)}dt = -\frac{dp}{\rho}\;\frac{1}{v} \\ \rho v\,dv &= -dp.\end{aligned}$$

From the continuity equation $\rho\,dv = -vd\rho$, we obtain

Equation:

$$\begin{array}{rcl} \rho v dv & = & -dp \\ (-vd\rho)v & = & -dp \\ v & = & \sqrt{\frac{dp}{d\rho}}. \end{array}$$

Consider a sound wave moving through air. During the process of compression and expansion of the gas, no heat is added or removed from the system. A process where heat is not added or removed from the system is known as an adiabatic system. Adiabatic processes are covered in detail in The First Law of Thermodynamics, but for now it is sufficient to say that for an adiabatic process,$pV^{\gamma} = \text{constant}$, where p is the pressure, V is the volume, and gamma (γ) is a constant that depends on the gas. For air, $\gamma = 1.40$. The density equals the number of moles times the molar mass divided by the volume, so the volume is equal to $V = \frac{nM}{\rho}$. The number of moles and the molar mass are constant and can be absorbed into the constant $p\left(\frac{1}{\rho}\right)^{\gamma} = \text{constant}$. Taking the natural logarithm of both sides yields $\ln p - \gamma \ln \rho = \text{constant}$. Differentiating with respect to the density, the equation becomes

Equation:

$$\begin{array}{rcl} \ln p - \gamma \ln \rho & = & \text{constant} \\ \frac{d}{d\rho}(\ln p - \gamma \ln \rho) & = & \frac{d}{d\rho}(\text{constant}) \\ \frac{1}{p}\frac{dp}{d\rho} - \frac{\gamma}{\rho} & = & 0 \\ \frac{dp}{d\rho} & = & \frac{\gamma p}{\rho}. \end{array}$$

If the air can be considered an ideal gas, we can use the ideal gas law:

Equation:

$$\begin{aligned} pV &= nRT = \tfrac{m}{M}RT \\ p &= \tfrac{m}{V}\,\tfrac{RT}{M} = \rho\tfrac{RT}{M}. \end{aligned}$$

Here *M* is the molar mass of air:

Equation:

$$\frac{dp}{d\rho} = \frac{\gamma p}{\rho} = \frac{\gamma\left(\rho\frac{RT}{M}\right)}{\rho} = \frac{\gamma RT}{M}.$$

Since the speed of sound is equal to $v = \sqrt{\frac{dp}{d\rho}}$, the speed is equal to

Equation:

$$v = \sqrt{\frac{\gamma\,RT}{M}}.$$

Note that the velocity is faster at higher temperatures and slower for heavier gases. For air, $\gamma = 1.4$, $M = 0.02897\frac{\text{kg}}{\text{mol}}$, and $R = 8.31\frac{\text{J}}{\text{mol·K}}$. If the temperature is $T_{\text{C}} = 20^{\circ}\text{C}\;(T = 293\,\text{K})$, the speed of sound is $v = 343\,\text{m/s}$.

The equation for the speed of sound in air $v = \sqrt{\frac{\gamma RT}{M}}$ can be simplified to give the equation for the speed of sound in air as a function of absolute temperature:

Equation:

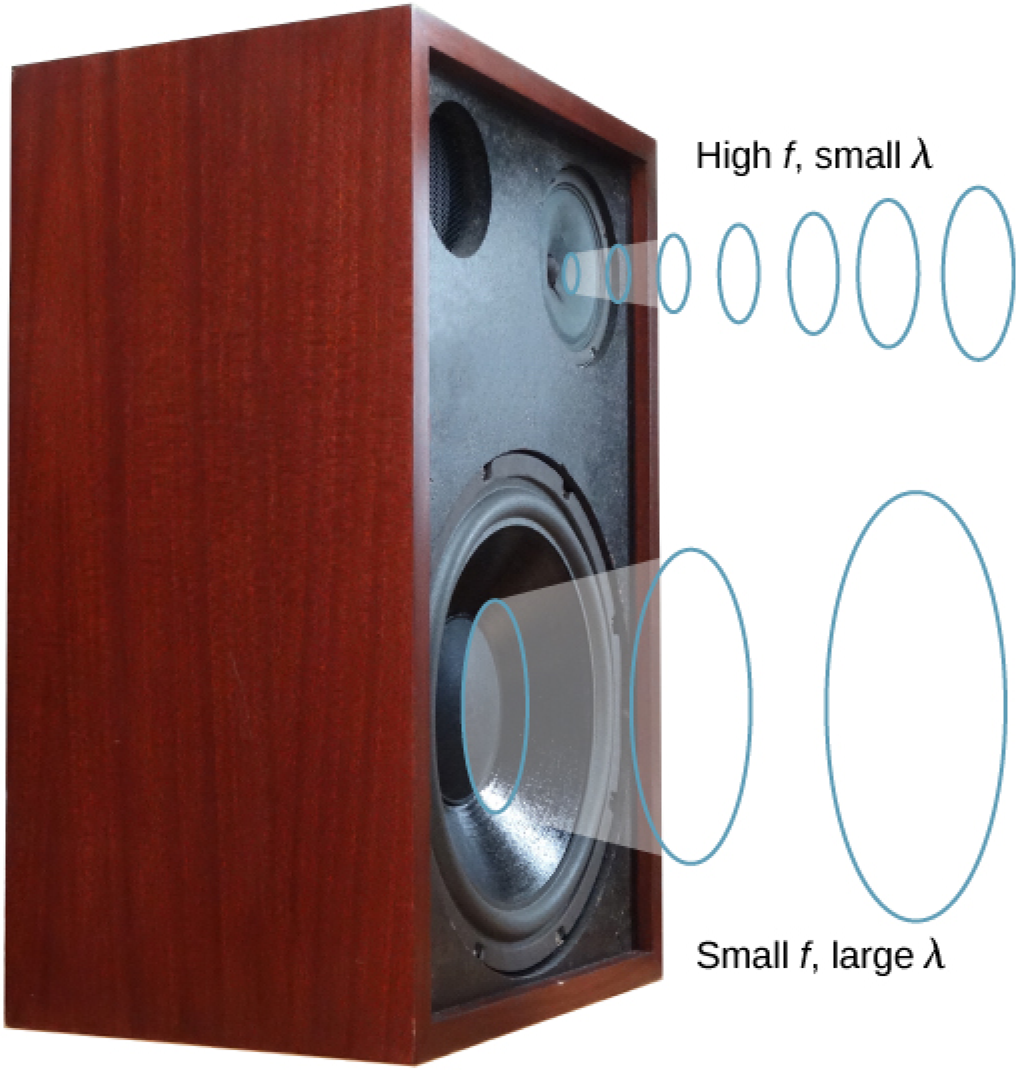
$$\begin{aligned} v &= \sqrt{\tfrac{\gamma RT}{M}} \\ &= \sqrt{\tfrac{\gamma RT}{M}\left(\tfrac{273\,\text{K}}{273\,\text{K}}\right)} = \sqrt{\tfrac{(273\,\text{K})\gamma R}{M}}\sqrt{\tfrac{T}{273\,\text{K}}} \\ &\approx 331\tfrac{\text{m}}{\text{s}}\sqrt{\tfrac{T}{273\,\text{K}}}. \end{aligned}$$

One of the more important properties of sound is that its speed is nearly independent of the frequency. This independence is certainly true in open air for sounds in the audible range. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, so all frequencies must travel at nearly the same speed. Recall that

Equation:

$$v = f\lambda.$$

In a given medium under fixed conditions, v is constant, so there is a relationship between f and λ; the higher the frequency, the smaller the wavelength ([link]).


High f, small λ
Small f, large λ

Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, whereas the higher-frequency sounds are emitted by the small speaker, called a tweeter. (credit: modification of work by Jane Whitney)

Example:

Calculating Wavelengths

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0°C air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v = f\lambda$.

Solution

1. Identify knowns. The value for *v* is given by
   Equation:
   $$v = (331\,\text{m/s})\sqrt{\frac{T}{273\,\text{K}}}.$$
2. Convert the temperature into kelvins and then enter the temperature into the equation
   Equation:
   $$v = (331\,\text{m/s})\sqrt{\frac{303\,\text{K}}{273\,\text{K}}} = 348.7\,\text{m/s}.$$
3. Solve the relationship between speed and wavelength for λ:
   Equation:
   $$\lambda = \frac{v}{f}.$$
4. Enter the speed and the minimum frequency to give the maximum wavelength:
   Equation:
   $$\lambda_{\text{max}} = \frac{348.7\,\text{m/s}}{20\text{ Hz}} = 17\,\text{m}.$$
5. Enter the speed and the maximum frequency to give the minimum wavelength:
   Equation:

$$\lambda_{\text{min}} = \frac{348.7\,\text{m/s}}{20{,}000\,\text{Hz}} = 0.017\,\text{m} = 1.7\,\text{cm.}$$

Significance
Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another, but the frequency usually remains the same. This is similar to the frequency of a wave on a string being equal to the frequency of the force oscillating the string. If v changes and f remains the same, then the wavelength λ must change. That is, because $v = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

Note:
Exercise:

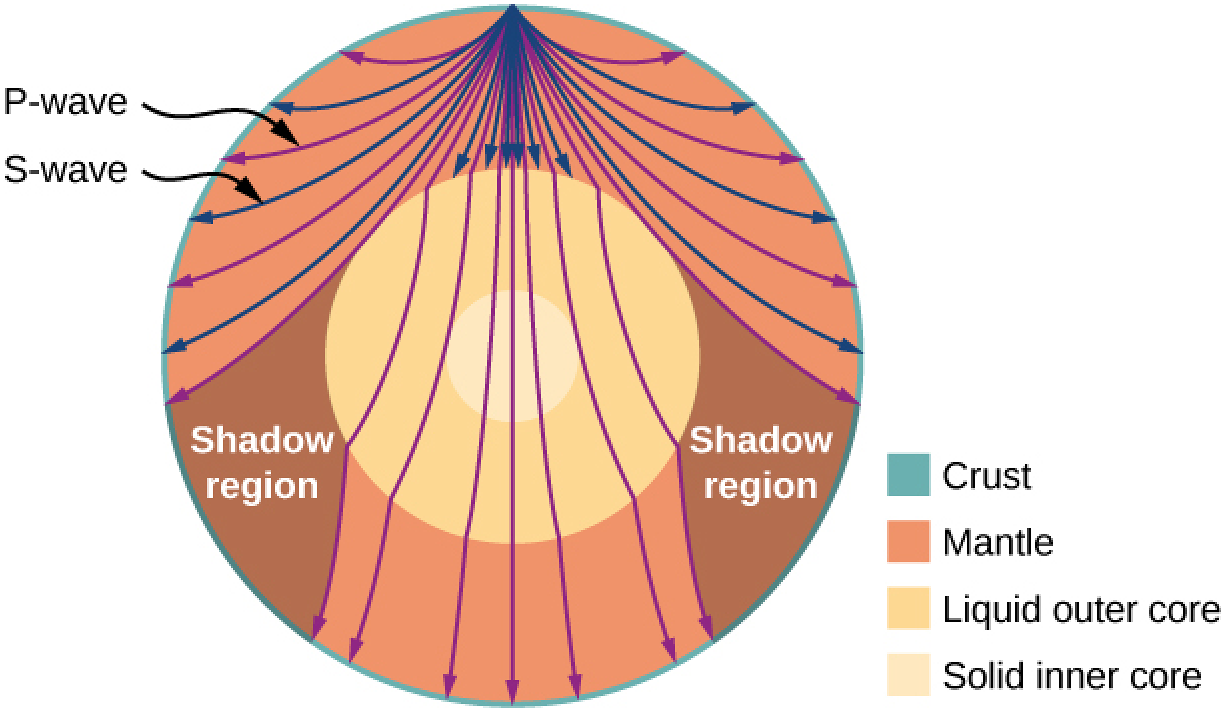
Problem:

Check Your Understanding Imagine you observe two firework shells explode. You hear the explosion of one as soon as you see it. However, you see the other shell for several milliseconds before you hear the explosion. Explain why this is so.

Solution:

Sound and light both travel at definite speeds, and the speed of sound is slower than the speed of light. The first shell is probably very close by, so the speed difference is not noticeable. The second shell is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Although sound waves in a fluid are longitudinal, sound waves in a solid travel both as longitudinal waves and transverse waves. Seismic waves, which are essentially sound waves in Earth's crust produced by earthquakes, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes produce both longitudinal and transverse waves, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both types of earthquake waves travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake. Because S-waves do not pass through the liquid core, two shadow regions are produced ([link]).



Earthquakes produce both longitudinal waves (P-waves) and transverse waves (S-waves), and these travel at different speeds. Both waves travel at different speeds in the different regions of Earth, but in general, P-waves travel faster than S-waves. S-waves cannot be supported by the liquid core, producing shadow regions.

As sound waves move away from a speaker, or away from the epicenter of an earthquake, their power per unit area decreases. This is why the sound is very loud near a speaker and becomes less loud as you move away from the speaker. This also explains why there can be an extreme amount of damage at the epicenter of an earthquake but only tremors are felt in areas far from the epicenter. The power per unit area is known as the intensity, and in the next section, we will discuss how the intensity depends on the distance from the source.

Summary

- The speed of sound depends on the medium and the state of the medium.
- In a fluid, because the absence of shear forces, sound waves are longitudinal. A solid can support both longitudinal and transverse sound waves.
- In air, the speed of sound is related to air temperature T by $v = 331\,\frac{\text{m}}{\text{s}}\sqrt{\frac{T_{\text{K}}}{273\,\text{K}}} = 331\,\frac{\text{m}}{\text{s}}\sqrt{1 + \frac{T_{\text{C}}}{273^{\circ}\text{C}}}$.
- v is the same for all frequencies and wavelengths of sound in air.

Conceptual Questions

Exercise:

Problem:

How do sound vibrations of atoms differ from thermal motion?

Exercise:

Problem:

When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

Solution:

The frequency does not change as the sound wave moves from one medium to another. Since the speed changes and the frequency does not, the wavelength must change. This is similar to the driving force of a harmonic oscillator or a wave on the string.

Exercise:

Problem:

A popular party trick is to inhale helium and speak in a high-frequency, funny voice. Explain this phenomenon.

Exercise:

Problem:

You may have used a sonic range finder in lab to measure the distance of an object using a clicking sound from a sound transducer. What is the principle used in this device?

Solution:

The transducer sends out a sound wave, which reflects off the object in question and measures the time it takes for the sound wave to return. Since the speed of sound is constant, the distance to the object can found by multiplying the velocity of sound by half the time interval measured.

Exercise:

Problem:

The sonic range finder discussed in the preceding question often needs to be calibrated. During the calibration, the software asks for the room temperature. Why do you suppose the room temperature is required?

Problems

Exercise:

Problem:

When poked by a spear, an operatic soprano lets out a 1200-Hz shriek. What is its wavelength if the speed of sound is 345 m/s?

Exercise:

Problem:

What frequency sound has a 0.10-m wavelength when the speed of sound is 340 m/s?

Solution:

$f=3400\text{ Hz}$

Exercise:

Problem:

Calculate the speed of sound on a day when a 1500-Hz frequency has a wavelength of 0.221 m.

Exercise:

Problem:

(a) What is the speed of sound in a medium where a 100-kHz frequency produces a 5.96-cm wavelength? (b) Which substance in [link] is this likely to be?

Solution:

a. $v = 5.96 \times 10^3 \, \mathrm{m/s}$; b. steel (from value in [link])

Exercise:

Problem:

Show that the speed of sound in $20.0^\circ\mathrm{C}$ air is $343 \, \mathrm{m/s}$, as claimed in the text.

Exercise:

Problem:

Air temperature in the Sahara Desert can reach $56.0^\circ\mathrm{C}$ (about $134^\circ\mathrm{F}$). What is the speed of sound in air at that temperature?

Solution:

$v = 363 \, \frac{\mathrm{m}}{\mathrm{s}}$

Exercise:

Problem:

Dolphins make sounds in air and water. What is the ratio of the wavelength of a sound in air to its wavelength in seawater? Assume air temperature is $20.0^\circ\mathrm{C}$.

Exercise:

Problem:

A sonar echo returns to a submarine 1.20 s after being emitted. What is the distance to the object creating the echo? (Assume that the submarine is in the ocean, not in fresh water.)

Solution:

$\Delta x = 924 \, \mathrm{m}$

Exercise:

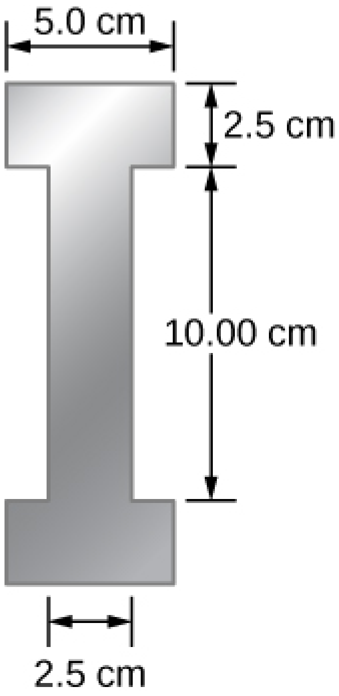
Problem:

(a) If a submarine’s sonar can measure echo times with a precision of 0.0100 s, what is the smallest difference in distances it can detect? (Assume that the submarine is in the ocean, not in fresh water.) (b) Discuss the limits this time resolution imposes on the ability of the sonar system to detect the size and shape of the object creating the echo.

Exercise:

Problem:

Ultrasonic sound waves are often used in methods of nondestructive testing. For example, this method can be used to find structural faults in a steel I-beams used in building. Consider a 10.00 meter long, steel I-beam with a cross-section shown below. The weight of the I-beam is 3846.50 N. What would be the speed of sound through in the I-beam? $(Y_{\text{steel}} = 200\ \text{GPa}, \beta_{\text{steel}} = 159\ \text{GPa})$.



Solution:

$V = 0.05\,\mathrm{m}^3$
$m = 392.5\,\mathrm{kg}$
$\rho = 7850\,\mathrm{kg/m}^3$
$v = 5047.54\,\mathrm{m/s}$

Exercise:

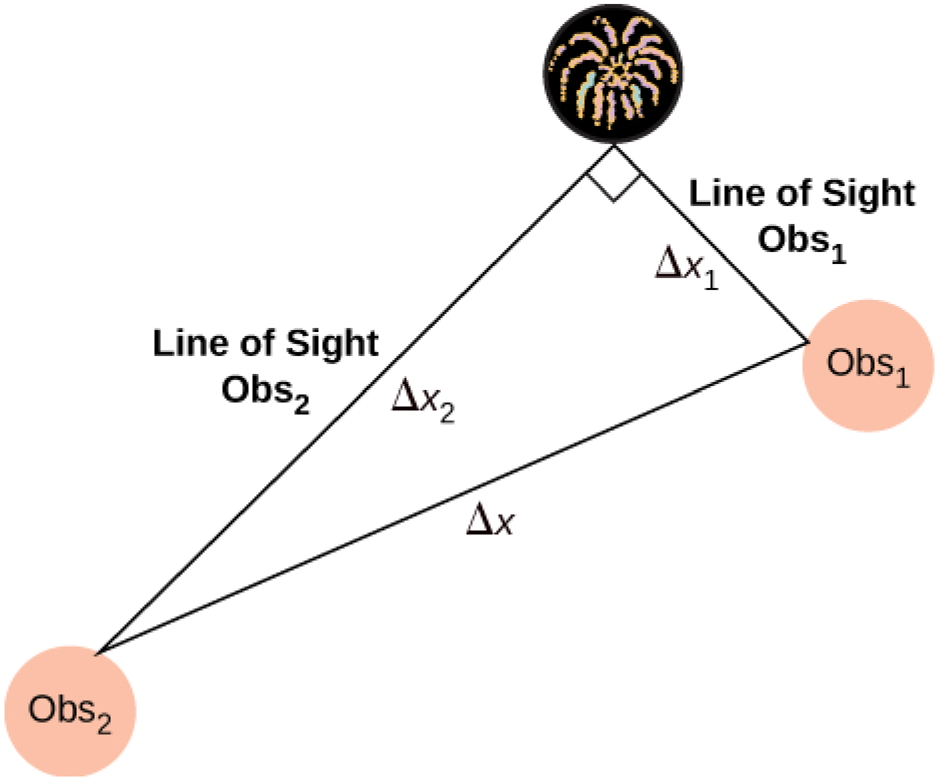
Problem:

A physicist at a fireworks display times the lag between seeing an explosion and hearing its sound, and finds it to be 0.400 s. (a) How far away is the explosion if air temperature is $24.0^\circ\mathrm{C}$ and if you neglect the time taken for light to reach the physicist? (b) Calculate the distance to the explosion taking the speed of light into account. Note that this distance is negligibly greater.

Exercise:

Problem:

During a 4th of July celebration, an M80 firework explodes on the ground, producing a bright flash and a loud bang. The air temperature of the night air is $T_\mathrm{F} = 90.00^\circ\mathrm{F}$. Two observers see the flash and hear the bang. The first observer notes the time between the flash and the bang as 1.00 second. The second observer notes the difference as 3.00 seconds. The line of sight between the two observers meet at a right angle as shown below. What is the distance Δx between the two observers?


Line of Sight
Obs$_1$
Δx_1
Line of Sight
Obs$_2$
Δx_2
Obs$_1$
Δx
Obs$_2$

Solution:

$$T_{\mathrm{C}} = 35^{\circ}\mathrm{C},\ \ v = 351.58\,\mathrm{m/s}$$
$$\Delta x_1 = 35.16\,\mathrm{m},\ \ \Delta x_2 = 52.74\,\mathrm{m}$$
$$\Delta x = 63.39\,\mathrm{m}$$

Exercise:

Problem:

The density of a sample of water is $\rho = 998.00\,\mathrm{kg/m}^3$ and the bulk modulus is $\beta = 2.15$ GPa. What is the speed of sound through the sample?

Exercise:

Problem:

Suppose a bat uses sound echoes to locate its insect prey, 3.00 m away. (See [link].) (a) Calculate the echo times for temperatures of 5.00°C and 35.0°C. (b) What percent uncertainty does this cause for the bat in locating the insect? (c) Discuss the significance of this uncertainty and whether it could cause difficulties for the bat. (In practice, the bat continues to use sound as it closes in, eliminating most of any difficulties imposed by this and other effects, such as motion of the prey.)

Solution:

a. $t_{5.00^\circ\text{C}} = 0.0180\text{ s},\quad t_{35.0^\circ\text{C}} = 0.0171\text{ s}$; b. $\%\text{ uncertainty} = 5.00\%$; c. This uncertainty could definitely cause difficulties for the bat, if it didn’t continue to use sound as it closed in on its prey. A 5% uncertainty could be the difference between catching the prey around the neck or around the chest, which means that it could miss grabbing its prey.

Sound Intensity

By the end of this section, you will be able to:

- Define the term intensity
- Explain the concept of sound intensity level
- Describe how the human ear translates sound

In a quiet forest, you can sometimes hear a single leaf fall to the ground. But when a passing motorist has his stereo turned up, you cannot even hear what the person next to you in your car is saying ([link]). We are all very familiar with the loudness of sounds and are aware that loudness is related to how energetically the source is vibrating. High noise exposure is hazardous to hearing, which is why it is important for people working in industrial settings to wear ear protection. The relevant physical quantity is sound intensity, a concept that is valid for all sounds whether or not they are in the audible range.

Noise on crowded roadways, like this one in Delhi, makes it hard to hear others unless they shout. (credit: “Lingaraj G J”/Flickr)

In Waves, we defined intensity as the power per unit area carried by a wave. Power is the rate at which energy is transferred by the wave. In equation form, intensity I is

Equation:

$$I = \frac{P}{A},$$

where P is the power through an area A. The SI unit for I is W/m^2. If we assume that the sound wave is spherical, and that no energy is lost to thermal processes, the energy of the sound wave is spread over a larger area as distance increases, so the intensity decreases. The area of a sphere is $A = 4\pi r^2$. As the wave spreads out from r_1 to r_2, the energy also spreads out over a larger area:

Equation:

$$\begin{aligned} P_1 &= P_2 \\ I_1 4\pi r_1^2 &= I_2 4\pi r_2^2; \end{aligned}$$

Note:

Equation:

$$I_2 = I_1 \left(\frac{r_1}{r_2}\right)^2.$$

The intensity decreases as the wave moves out from the source. In an inverse square relationship, such as the intensity, when you double the distance, the intensity decreases to one quarter,

Equation:

$$I_2 = I_1 \left(\frac{r_1}{r_2}\right)^2 = I_1 \left(\frac{r_1}{2r_1}\right)^2 = \frac{1}{4} I_1.$$

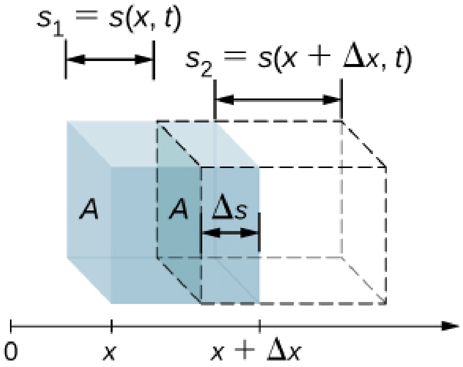
Generally, when considering the intensity of a sound wave, we take the intensity to be the time-averaged value of the power, denoted by $\langle P \rangle$, divided by the area,

Note:

Equation:

$$I = \frac{\langle P \rangle}{A}.$$

The intensity of a sound wave is proportional to the change in the pressure squared and inversely proportional to the density and the speed. Consider a parcel of a medium initially undisturbed and then influenced by a sound wave at time *t*, as shown in [link].



An undisturbed parcel of a medium with a volume $V = A\Delta x$ shown in blue. A sound wave moves through the medium at time *t*, and the parcel is displaced and expands, as shown by dotted lines. The change in volume is $\Delta V = A\Delta s = A\left(s_2 - s_1\right)$, where s_1 is the displacement of the leading edge of the parcel and s_2 is the displacement of the trailing edge of the parcel. In the figure, $s_2 > s_1$ and the parcel expands, but the parcel can either expand or compress $\left(s_2 < s_1\right)$, depending on which part of the sound wave (compression or rarefaction) is moving through the parcel.

As the sound wave moves through the parcel, the parcel is displaced and may expand or contract. If $s_2 > s_1$, the volume has increased and the pressure decreases. If $s_2 < s_1$, the volume has decreased and the pressure increases. The change in the volume is

Equation:

$$\Delta V = A\Delta s = A\left(s_2 - s_1\right) = A\left(s\left(x + \Delta x, t\right) - s\left(x, t\right)\right).$$

The fractional change in the volume is the change in volume divided by the original volume:

Equation:

$$\frac{dV}{V} = \lim_{\Delta x \to 0} \frac{A\left[s\left(x+\Delta x, t\right) - s\left(x, t\right)\right]}{A\Delta x} = \frac{\partial s\left(x, t\right)}{\partial x}.$$

The fractional change in volume is related to the pressure fluctuation by the bulk modulus $\beta = -\frac{\Delta p(x,t)}{dV/V}$. Recall that the minus sign is required because the volume is *inversely* related to the pressure. (We use lowercase *p* for pressure to distinguish it from power, denoted by *P*.) The change in pressure is therefore $\Delta p\left(x, t\right) = -\beta \frac{dV}{V} = -\beta \frac{\partial s(x,t)}{\partial x}$. If the sound wave is sinusoidal, then the displacement as shown in [link] is $s\left(x, t\right) = s_{\text{max}}\text{cos}\left(kx \mp \omega t + \phi\right)$ and the pressure is found to be

Equation:

$$\Delta p\left(x, t\right) = -\beta \frac{dV}{V} = -\beta \frac{\partial s\left(x, t\right)}{\partial x} = \beta k s_{\text{max}}\text{sin}\left(kx - \omega t + \phi\right) = \Delta p_{\text{max}}\text{sin}\left(kx - \omega t + \phi\right).$$

The intensity of the sound wave is the power per unit area, and the power is the force times the velocity, $I = \frac{P}{A} = \frac{Fv}{A} = pv$. Here, the velocity is the velocity of the oscillations of the medium, and not the velocity of the sound wave. The velocity of the medium is the time rate of change in the displacement:

Equation:

$$v\left(x, t\right) = \frac{\partial}{\partial y} s\left(x, t\right) = \frac{\partial}{\partial y}\left(s_{\text{max}}\text{cos}\left(kx - \omega t + \phi\right)\right) = s_{\text{max}}\omega\,\text{sin}\left(kx - \omega t + \phi\right).$$

Thus, the intensity becomes

Equation:

$$\begin{aligned} I &= \Delta p\left(x, t\right) v\left(x, t\right) \\ &= \beta k s_{\text{max}}\text{sin}\left(kx - \omega t + \phi\right)\left[s_{\text{max}}\omega\,\text{sin}\left(kx - \omega t + \phi\right)\right] \\ &= \beta k \omega s_{\text{max}}^2 \text{sin}^2\left(kx - \omega t + \phi\right). \end{aligned}$$

To find the time-averaged intensity over one period $T = \frac{2\pi}{\omega}$ for a position *x*, we integrate over the period, $I = \frac{\beta k \omega s_{\text{max}}^2}{2}$. Using $\Delta p_{\text{max}} = \beta k s_{\text{max}}$, $v = \sqrt{\frac{\beta}{\rho}}$, and $v = \frac{\omega}{k}$, we obtain

Equation:

$$I = \frac{\beta k \omega s_{\text{max}}^2}{2} = \frac{\beta^2 k^2 \omega s_{\text{max}}^2}{2\beta k} = \frac{\omega \left(\Delta p_{\text{max}}\right)^2}{2\left(\rho v^2\right)k} = \frac{v \left(\Delta p_{\text{max}}\right)^2}{2\left(\rho v^2\right)} = \frac{\left(\Delta p_{\text{max}}\right)^2}{2\rho v}.$$

That is, the intensity of a sound wave is related to its amplitude squared by

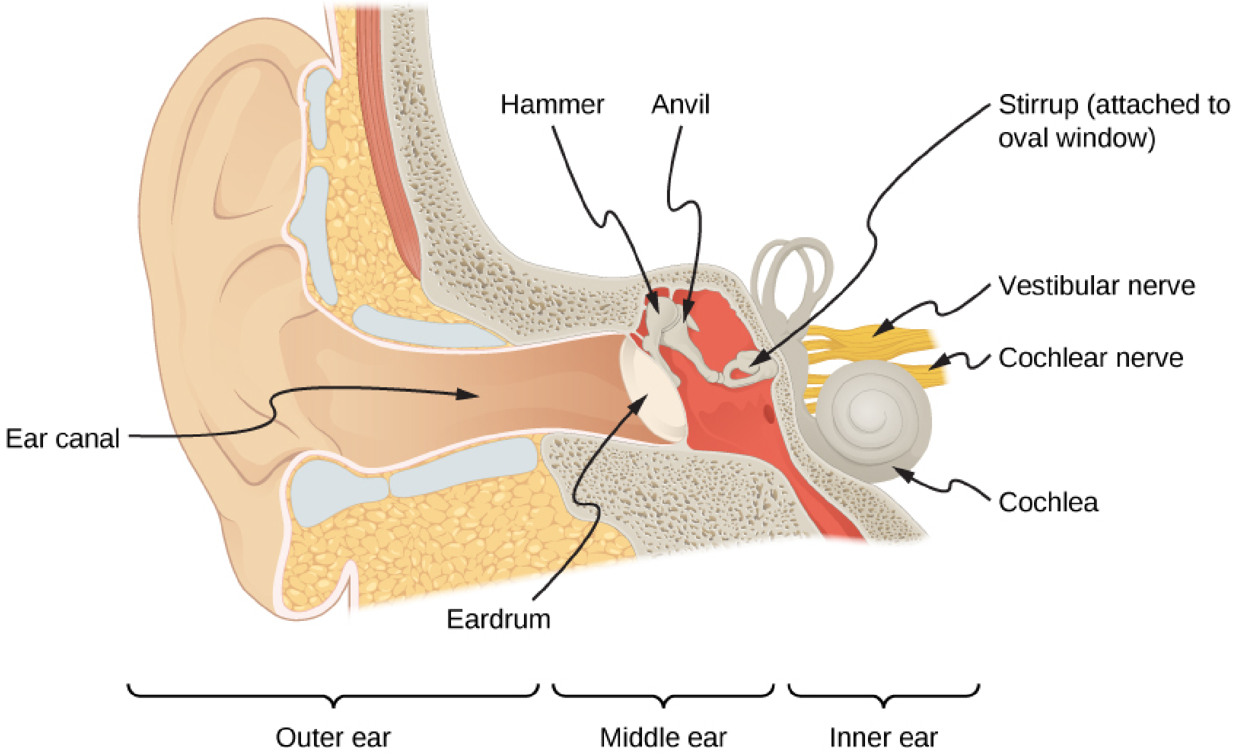
Note:
Equation:

$$I = \frac{(\Delta p_{\text{max}})^2}{2\rho v}.$$

Here, Δp_{max} is the pressure variation or pressure amplitude in units of pascals (Pa) or N/m^2. The energy (as kinetic energy $\frac{1}{2}mv^2$) of an oscillating element of air due to a traveling sound wave is proportional to its amplitude squared. In this equation, ρ is the density of the material in which the sound wave travels, in units of kg/m^3, and v is the speed of sound in the medium, in units of m/s. The pressure variation is proportional to the amplitude of the oscillation, so I varies as $(\Delta p)^2$. This relationship is consistent with the fact that the sound wave is produced by some vibration; the greater its pressure amplitude, the more the air is compressed in the sound it creates.

Human Hearing and Sound Intensity Levels

As stated earlier in this chapter, hearing is the perception of sound. The hearing mechanism involves some interesting physics. The sound wave that impinges upon our ear is a pressure wave. The ear is a **transducer** that converts sound waves into electrical nerve impulses in a manner much more sophisticated than, but analogous to, a microphone. [link] shows the anatomy of the ear.


Hammer
Anvil
Stirrup (attached to oval window)
Vestibular nerve
Cochlear nerve
Cochlea
Ear canal
Eardrum
Outer ear
Middle ear
Inner ear

The anatomy of the human ear.

The outer ear, or ear canal, carries sound to the recessed, protected eardrum. The air column in the ear canal resonates and is partially responsible for the sensitivity of the ear to sounds in the 2000–5000-Hz range. The middle ear converts sound into mechanical vibrations and applies these vibrations to the cochlea.

Note:
Watch this video for a more detailed discussion of the workings of the human ear.

The range of intensities that the human ear can hear depends on the frequency of the sound, but, in general, the range is quite large. The minimum threshold intensity that can be heard is $I_0 = 10^{-12}\ \mathrm{W/m^2}$. Pain is experienced at intensities of $I_{\text{pain}} = 1\ \mathrm{W/m^2}$. Measurements of sound intensity (in units of $\mathrm{W/m^2}$) are very cumbersome due to this large range in values. For this reason, as well as for other reasons, the concept of sound intensity level was proposed.

The **sound intensity level** β of a sound, measured in decibels, having an intensity I in watts per meter squared, is defined as

Note:
Equation:

$$\beta(\text{dB}) = 10 \log_{10}\left(\frac{I}{I_0}\right),$$

where $I_0 = 10^{-12}\ \text{W/m}^2$ is a reference intensity, corresponding to the threshold intensity of sound that a person with normal hearing can perceive at a frequency of 1.00 kHz. It is more common to consider sound intensity levels in dB than in W/m^2. How human ears perceive sound can be more accurately described by the logarithm of the intensity rather than directly by the intensity. Because β is defined in terms of a ratio, it is a unitless quantity, telling you the *level* of the sound relative to a fixed standard ($10^{-12}\ \text{W/m}^2$). The units of decibels (dB) are used to indicate this ratio is multiplied by 10 in its definition. The bel, upon which the decibel is based, is named for Alexander Graham Bell, the inventor of the telephone.

The decibel level of a sound having the threshold intensity of $10^{-12}\ \text{W/m}^2$ is $\beta = 0\ \text{dB}$, because $\log_{10} 1 = 0$. [link] gives levels in decibels and intensities in watts per meter squared for some familiar sounds. The ear is sensitive to as little as a trillionth of a watt per meter squared—even more impressive when you realize that the area of the eardrum is only about $1\ \text{cm}^2$, so that only $10^{-16}\ \text{W}$ falls on it at the threshold of hearing. Air molecules in a sound wave of this intensity vibrate over a distance of less than one molecular diameter, and the gauge pressures involved are less than $10^{-9}\ \text{atm}$.

Sound intensity level β (dB)	**Intensity *I*** $\left(\text{W/m}^2\right)$	**Example/effect**
0	1×10^{-12}	Threshold of hearing at 1000 Hz
10	1×10^{-11}	Rustle of leaves
20	1×10^{-10}	Whisper at 1-m distance
30	1×10^{-9}	Quiet home
40	1×10^{-8}	Average home

Sound intensity level β (dB)	**Intensity *I* $\left(\text{W/m}^2\right)$**	**Example/effect**
50	$1\ \times\ 10^{-7}$	Average office, soft music
60	$1\ \times\ 10^{-6}$	Normal conversation
70	$1\ \times\ 10^{-5}$	Noisy office, busy traffic
80	$1\ \times\ 10^{-4}$	Loud radio, classroom lecture
90	$1\ \times\ 10^{-3}$	Inside a heavy truck; damage from prolonged exposure[footnote] Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.
100	$1\ \times\ 10^{-2}$	Noisy factory, siren at 30 m; damage from 8 h per day exposure
110	$1\ \times\ 10^{-1}$	Damage from 30 min per day exposure
120	1	Loud rock concert; pneumatic chipper at 2 m; threshold of pain
140	$1\ \times\ 10^{2}$	Jet airplane at 30 m; severe pain, damage in seconds
160	$1\ \times\ 10^{4}$	Bursting of eardrums

Sound Intensity Levels and Intensities[1] Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.

An observation readily verified by examining [link] or by using [link] is that each factor of 10 in intensity corresponds to 10 dB. For example, a 90-dB sound compared with a 60-dB sound is 30 dB greater, or three factors of 10 (that is, 10^3 times) as intense. Another example is that if one sound is 10^7 as intense as another, it is 70 dB higher ([link]).

I_2/I_1	$\beta_2 - \beta_1$
2.0	3.0 dB
5.0	7.0 dB
10.0	10.0 dB
100.0	20.0 dB
1000.0	30.0 dB

Ratios of Intensities and Corresponding Differences in Sound Intensity Levels

Example:

Calculating Sound Intensity Levels

Calculate the sound intensity level in decibels for a sound wave traveling in air at 0°C and having a pressure amplitude of 0.656 Pa.

Strategy

We are given Δp, so we can calculate I using the equation $I = \frac{(\Delta p)^2}{2\rho v_\text{w}}$. Using I, we can calculate β straight from its definition in $\beta(dB) = 10\log_{10}\left(\frac{I}{I_0}\right)$.

Solution

1. Identify knowns:
   Sound travels at 331 m/s in air at 0°C.
   Air has a density of $1.29\ \text{kg/m}^3$ at atmospheric pressure and 0°C.
2. Enter these values and the pressure amplitude into $I = \frac{(\Delta p)^2}{2\rho v}$.
   Equation:
   $$I = \frac{(\Delta p)^2}{2\rho v} = \frac{(0.656\ \text{Pa})^2}{2(1.29\ \text{kg/m}^3)(331\ \text{m/s})} = 5.04\ \times\ 10^{-4}\ \text{W/m}^2.$$
3. Enter the value for I and the known value for I_0 into $\beta(\text{dB}) = 10\log_{10}(I/I_0)$. Calculate to find the sound intensity level in decibels:
   Equation:
   $$10\log_{10}(5.04\ \times\ 10^8) = 10(8.70)\text{dB} = 87\ \text{dB}.$$

Significance

This 87-dB sound has an intensity five times as great as an 80-dB sound. So a factor of five in intensity corresponds to a difference of 7 dB in sound intensity level. This value is true for any intensities differing by a factor of five.

Example:

Changing Intensity Levels of a Sound

Show that if one sound is twice as intense as another, it has a sound level about 3 dB higher.

Strategy

We are given that the ratio of two intensities is 2 to 1, and are then asked to find the difference in their sound levels in decibels. We can solve this problem by using of the properties of logarithms.

Solution

1. Identify knowns:
   The ratio of the two intensities is 2 to 1, or
   Equation:
   $$\frac{I_2}{I_1} = 2.00.$$
   We wish to show that the difference in sound levels is about 3 dB. That is, we want to show:
   Equation:
   $$\beta_2 - \beta_1 = 3\text{ dB}.$$
   Note that
   Equation:
   $$\log_{10}b - \log_{10}a = \log_{10}\left(\frac{b}{a}\right).$$
2. Use the definition of β to obtain
   Equation:
   $$\beta_2 - \beta_1 = 10\log_{10}\left(\frac{I_2}{I_1}\right) = 10\log_{10}2.00 = 10(0.301)\text{ dB}.$$
   Thus,
   Equation:
   $$\beta_2 - \beta_1 = 3.01\text{ dB}.$$

Significance

This means that the two sound intensity levels differ by 3.01 dB, or about 3 dB, as advertised. Note that because only the ratio I_2/I_1 is given (and not the actual intensities), this result is true for any intensities that differ by a factor of two. For example, a 56.0-dB sound is twice as intense as a 53.0-dB sound, a 97.0-dB sound is half as intense as a 100-dB sound, and so on.

Note:
Exercise:

Problem:

Check Your Understanding Identify common sounds at the levels of 10 dB, 50 dB, and 100 dB.

Solution:

10 dB: rustle of leaves; 50 dB: average office; 100 dB: noisy factory

Another decibel scale is also in use, called the **sound pressure level**, based on the ratio of the pressure amplitude to a reference pressure. This scale is used particularly in applications where sound travels in water. It is beyond the scope of this text to treat this scale because it is not commonly used for sounds in air, but it is important to note that very different decibel levels may be encountered when sound pressure levels are quoted.

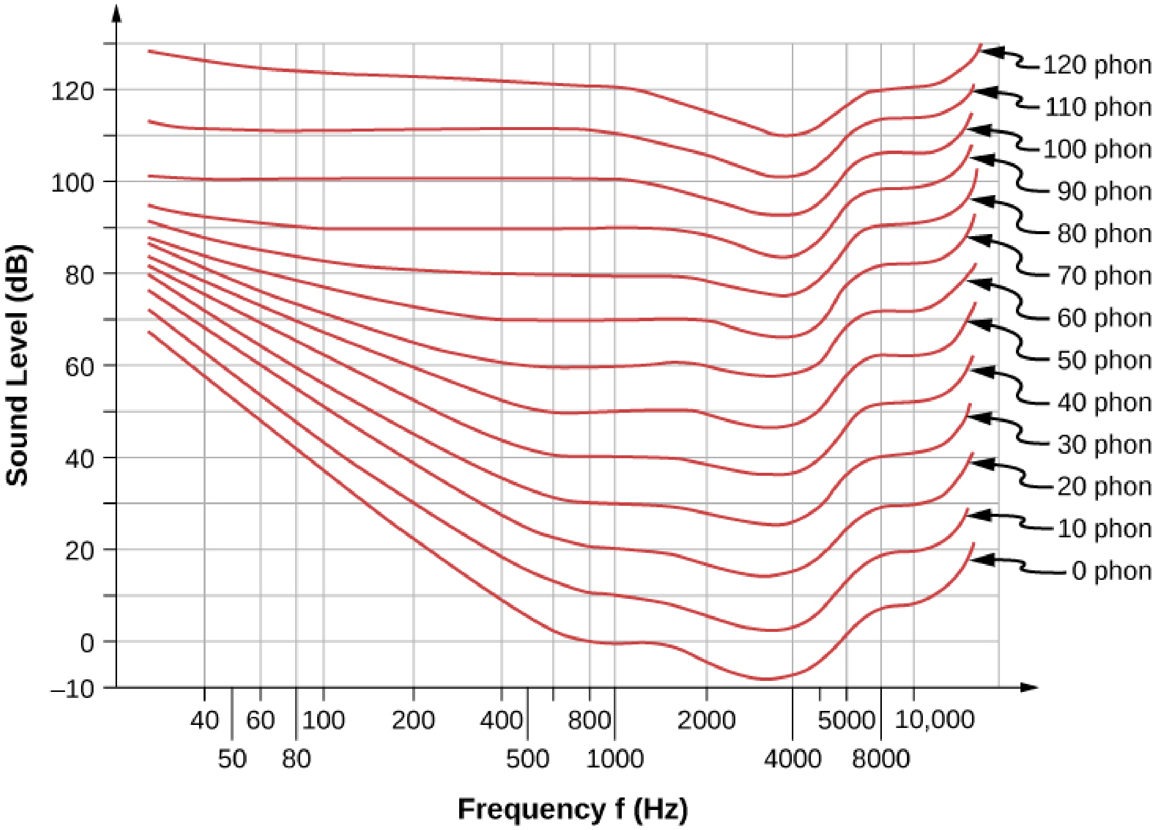
Hearing and Pitch

The human ear has a tremendous range and sensitivity. It can give us a wealth of simple information—such as pitch, loudness, and direction.

The perception of frequency is called **pitch**. Typically, humans have excellent relative pitch and can discriminate between two sounds if their frequencies differ by 0.3% or more. For example, 500.0 and 501.5 Hz are noticeably different. Musical **notes** are sounds of a particular frequency that can be produced by most instruments and in Western music have particular names, such as A-sharp, C, or E-flat.

The perception of intensity is called **loudness**. At a given frequency, it is possible to discern differences of about 1 dB, and a change of 3 dB is easily noticed. But loudness is not related to intensity alone. Frequency has a major effect on how loud a sound seems. Sounds near the high- and low-frequency extremes of the hearing range seem even less loud, because the ear is less sensitive at those frequencies. When a violin plays middle C, there is no mistaking it for a piano playing the same note. The reason is that each instrument produces a distinctive set of frequencies and intensities. We call our perception of these combinations of frequencies and intensities tone quality or, more commonly, the **timbre** of the sound. Timbre is the shape of the wave that arises from the many reflections, resonances, and superposition in an instrument.

A unit called a **phon** is used to express loudness numerically. Phons differ from decibels because the phon is a unit of loudness perception, whereas the decibel is a unit of physical intensity. [link] shows the relationship of loudness to intensity (or intensity level) and frequency for persons with normal hearing. The curved lines are equal-loudness curves. Each curve is labeled with its loudness in phons. Any sound along a given curve is perceived as equally loud by the average person. The curves were determined by having large numbers of people compare the loudness of sounds at different frequencies and sound intensity levels. At a frequency of 1000 Hz, phons are taken to be numerically equal to decibels.



The relationship of loudness in phons to intensity level (in decibels) and intensity (in watts per meter squared) for persons with normal hearing. The curved lines are equal-loudness curves—all sounds on a given curve are perceived as equally loud. Phons and decibels are defined to be the same at 1000 Hz.

Example:

Measuring Loudness

(a) What is the loudness in phons of a 100-Hz sound that has an intensity level of 80 dB? (b) What is the intensity level in decibels of a 4000-Hz sound having a loudness of 70 phons? (c) At what intensity level will an 8000-Hz sound have the same loudness as a 200-Hz sound at 60 dB?

Strategy

The graph in [link] should be referenced to solve this example. To find the loudness of a given sound, you must know its frequency and intensity level, locate that point on the square grid, and then interpolate between loudness curves to get the loudness in phons. Once that point is located, the intensity level can be determined from the vertical axis.

Solution

1. Identify knowns: The square grid of the graph relating phons and decibels is a plot of intensity level versus frequency—both physical quantities: 100 Hz at 80 dB lies halfway between the curves marked 70 and 80 phons.
   Find the loudness: 75 phons.
2. Identify knowns: Values are given to be 4000 Hz at 70 phons.
   Follow the 70-phon curve until it reaches 4000 Hz. At that point, it is below the 70 dB line at about 67 dB.
   Find the intensity level: 67 dB.
3. Locate the point for a 200 Hz and 60 dB sound.
   Find the loudness: This point lies just slightly above the 50-phon curve, and so its loudness is 51 phons.
   Look for the 51-phon level is at 8000 Hz: 63 dB.

Significance
These answers, like all information extracted from [link], have uncertainties of several phons or several decibels, partly due to difficulties in interpolation, but mostly related to uncertainties in the equal-loudness curves.

Note:
Exercise:

Problem:

Check Your Understanding Describe how amplitude is related to the loudness of a sound.

Solution:

Amplitude is directly proportional to the experience of loudness. As amplitude increases, loudness increases.

In this section, we discussed the characteristics of sound and how we hear, but how are the sounds we hear produced? Interesting sources of sound are musical instruments and the human voice, and we will discuss these sources. But before we can understand how musical instruments produce sound, we need to look at the basic mechanisms behind these instruments. The theories behind the mechanisms used by musical instruments involve interference, superposition, and standing waves, which we discuss in the next section.

Summary

- Intensity $I = P/A$ is the same for a sound wave as was defined for all waves, where P is the power crossing area A. The SI unit for I is watts per meter squared. The intensity of a sound wave is also related to the pressure amplitude Δp:

Equation:

$$I = \frac{(\Delta p)^2}{2\,\rho v},$$

where ρ is the density of the medium in which the sound wave travels and v_w is the speed of sound in the medium.

- Sound intensity level in units of decibels (dB) is

  Equation:

$$\beta(\text{dB}) = 10\log_{10}\left(\frac{I}{I_0}\right),$$

  where $I_0 = 10^{-12}\ \text{W/m}^2$ is the threshold intensity of hearing.
- The perception of frequency is pitch. The perception of intensity is loudness and loudness has units of phons.

Conceptual Questions

Exercise:

Problem:

Six members of a synchronized swim team wear earplugs to protect themselves against water pressure at depths, but they can still hear the music and perform the combinations in the water perfectly. One day, they were asked to leave the pool so the dive team could practice a few dives, and they tried to practice on a mat, but seemed to have a lot more difficulty. Why might this be?

Solution:

The ear plugs reduce the intensity of the sound both in water and on land, but Navy researchers have found that sound under water is heard through vibrations mastoid, which is the bone behind the ear.

Exercise:

Problem:

A community is concerned about a plan to bring train service to their downtown from the town’s outskirts. The current sound intensity level, even though the rail yard is blocks away, is 70 dB downtown. The mayor assures the public that there will be a difference of only 30 dB in sound in the downtown area. Should the townspeople be concerned? Why?

Problems

Exercise:

Problem: What is the intensity in watts per meter squared of a 85.0-dB sound?

Exercise:

Problem:

The warning tag on a lawn mower states that it produces noise at a level of 91.0 dB. What is this in watts per meter squared?

Solution:

$1.26 \times 10^{-3}\,\mathrm{W/m^2}$

Exercise:

Problem:

A sound wave traveling in air has a pressure amplitude of 0.5 Pa. What is the intensity of the wave?

Exercise:

Problem: What intensity level does the sound in the preceding problem correspond to?

Solution:

85 dB

Exercise:

Problem:

What sound intensity level in dB is produced by earphones that create an intensity of $4.00 \times 10^{-2}\,\mathrm{W/m^2}$?

Exercise:

Problem:

What is the decibel level of a sound that is twice as intense as a 90.0-dB sound? (b) What is the decibel level of a sound that is one-fifth as intense as a 90.0-dB sound?

Solution:

a. 93 dB; b. 83 dB

Exercise:

Problem:

What is the intensity of a sound that has a level 7.00 dB lower than a $4.00 \times 10^{-9}\text{-W/m}^2$ sound? (b) What is the intensity of a sound that is 3.00 dB higher than a $4.00 \times 10^{-9}\text{-W/m}^2$ sound?

Exercise:

Problem:

People with good hearing can perceive sounds as low as −8.00 dB at a frequency of 3000 Hz. What is the intensity of this sound in watts per meter squared?

Solution:

$1.58 \times 10^{-13}\ \text{W/m}^2$

Exercise:

Problem:

If a large housefly 3.0 m away from you makes a noise of 40.0 dB, what is the noise level of 1000 flies at that distance, assuming interference has a negligible effect?

Exercise:

Problem:

Ten cars in a circle at a boom box competition produce a 120-dB sound intensity level at the center of the circle. What is the average sound intensity level produced there by each stereo, assuming interference effects can be neglected?

Solution:

A decrease of a factor of 10 in intensity corresponds to a reduction of 10 dB in sound level: $120\ \text{dB} - 10\ \text{dB} = 110\ \text{dB}$.

Exercise:

Problem:

The amplitude of a sound wave is measured in terms of its maximum gauge pressure. By what factor does the amplitude of a sound wave increase if the sound intensity level goes up by 40.0 dB?

Exercise:

Problem:

If a sound intensity level of 0 dB at 1000 Hz corresponds to a maximum gauge pressure (sound amplitude) of $10^{-9}\ \text{atm}$, what is the maximum gauge pressure in a 60-dB sound? What is the maximum gauge pressure in a 120-dB sound?

Solution:

We know that 60 dB corresponds to a factor of 10^6 increase in intensity. Therefore, $I \propto X^2 \Rightarrow \frac{I_2}{I_1} = \left(\frac{X_2}{X_1}\right)^2$, so that $X_2 = 10^{-6}\,\text{atm}$.

120 dB corresponds to a factor of 10^{12} increase$\Rightarrow 10^{-9}\,\text{atm}(10^{12})^{1/2} = 10^{-3}\,\text{atm}$.

Exercise:

Problem:

An 8-hour exposure to a sound intensity level of 90.0 dB may cause hearing damage. What energy in joules falls on a 0.800-cm-diameter eardrum so exposed?

Exercise:

Problem:

Sound is more effectively transmitted into a stethoscope by direct contact rather than through the air, and it is further intensified by being concentrated on the smaller area of the eardrum. It is reasonable to assume that sound is transmitted into a stethoscope 100 times as effectively compared with transmission though the air. What, then, is the gain in decibels produced by a stethoscope that has a sound gathering area of $15.0\,\text{cm}^2$, and concentrates the sound onto two eardrums with a total area of $0.900\,\text{cm}^2$ with an efficiency of 40.0%?

Solution:

28.2 dB

Exercise:

Problem:

Loudspeakers can produce intense sounds with surprisingly small energy input in spite of their low efficiencies. Calculate the power input needed to produce a 90.0-dB sound intensity level for a 12.0-cm-diameter speaker that has an efficiency of 1.00%. (This value is the sound intensity level right at the speaker.)

Exercise:

Problem:

The factor of 10^{-12} in the range of intensities to which the ear can respond, from threshold to that causing damage after brief exposure, is truly remarkable. If you could measure distances over the same range with a single instrument and the smallest distance you could measure was 1 mm, what would the largest be?

Solution:

$1 \times 10^6\,\text{km}$

Exercise:

Problem:

What are the closest frequencies to 500 Hz that an average person can clearly distinguish as being different in frequency from 500 Hz? The sounds are not present simultaneously.

Exercise:

Problem:

Can you tell that your roommate turned up the sound on the TV if its average sound intensity level goes from 70 to 73 dB?

Solution:

$73\text{ dB}-70\text{ dB}=3\text{ dB}$; Such a change in sound level is easily noticed.

Exercise:

Problem:

If a woman needs an amplification of $5.0\ \times\ 10^{5}$ times the threshold intensity to enable her to hear at all frequencies, what is her overall hearing loss in dB? Note that smaller amplification is appropriate for more intense sounds to avoid further damage to her hearing from levels above 90 dB.

Exercise:

Problem:

A person has a hearing threshold 10 dB above normal at 100 Hz and 50 dB above normal at 4000 Hz. How much more intense must a 100-Hz tone be than a 4000-Hz tone if they are both barely audible to this person?

Solution:

2.5; The 100-Hz tone must be 2.5 times more intense than the 4000-Hz sound to be audible by this person.

Glossary

loudness
perception of sound intensity

notes
basic unit of music with specific names, combined to generate tunes

phon
numerical unit of loudness

pitch

perception of the frequency of a sound

sound intensity level
unitless quantity telling you the level of the sound relative to a fixed standard

sound pressure level
ratio of the pressure amplitude to a reference pressure

timbre
number and relative intensity of multiple sound frequencies

transducer
device that converts energy of a signal into measurable energy form, for example, a microphone converts sound waves into an electrical signal

Normal Modes of a Standing Sound Wave

By the end of this section, you will be able to:

- Explain the mechanism behind sound-reducing headphones
- Describe resonance in a tube closed at one end and open at the other end
- Describe resonance in a tube open at both ends

Interference is the hallmark of waves, all of which exhibit constructive and destructive interference exactly analogous to that seen for water waves. In fact, one way to prove something “is a wave” is to observe interference effects. Since sound is a wave, we expect it to exhibit interference.

Interference of Sound Waves

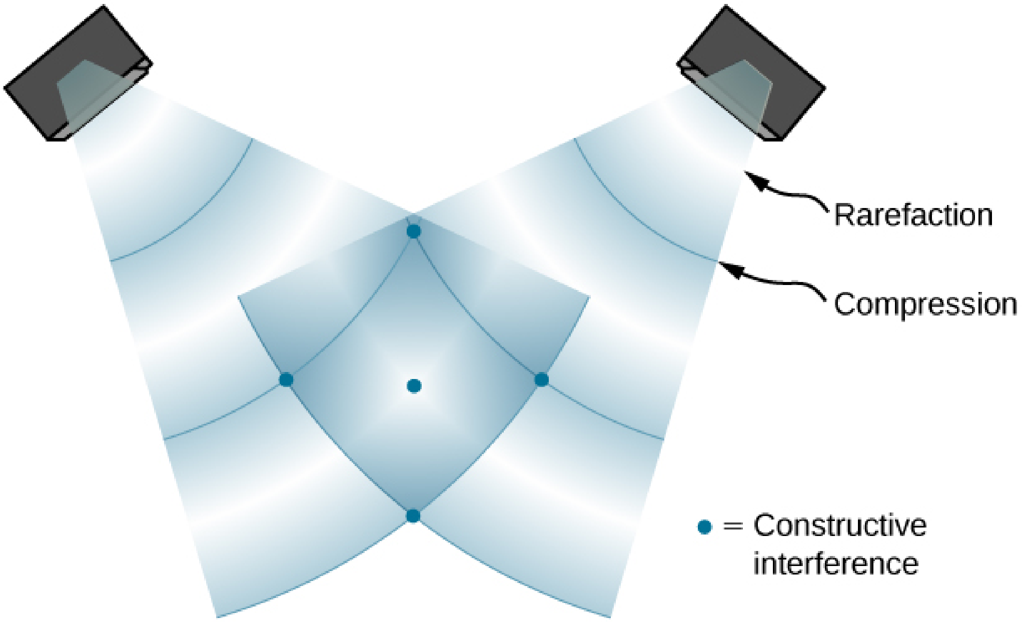
In Waves, we discussed the interference of wave functions that differ only in a phase shift. We found that the wave function resulting from the superposition of $y_1(x,t) = A\sin(kx-\omega t+\phi)$ and $y_2(x,t) = A\sin(kx-\omega t)$ is

Equation:

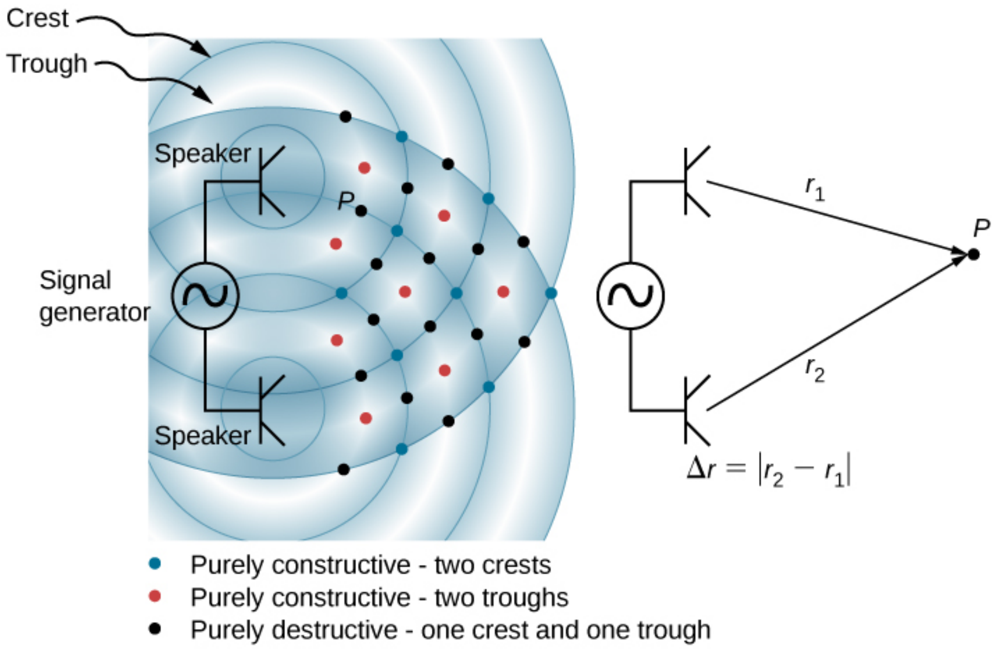
$$y(x,t) = \left[2A\cos\left(\frac{\phi}{2}\right)\right]\sin\left(kx-\omega t+\frac{\phi}{2}\right).$$

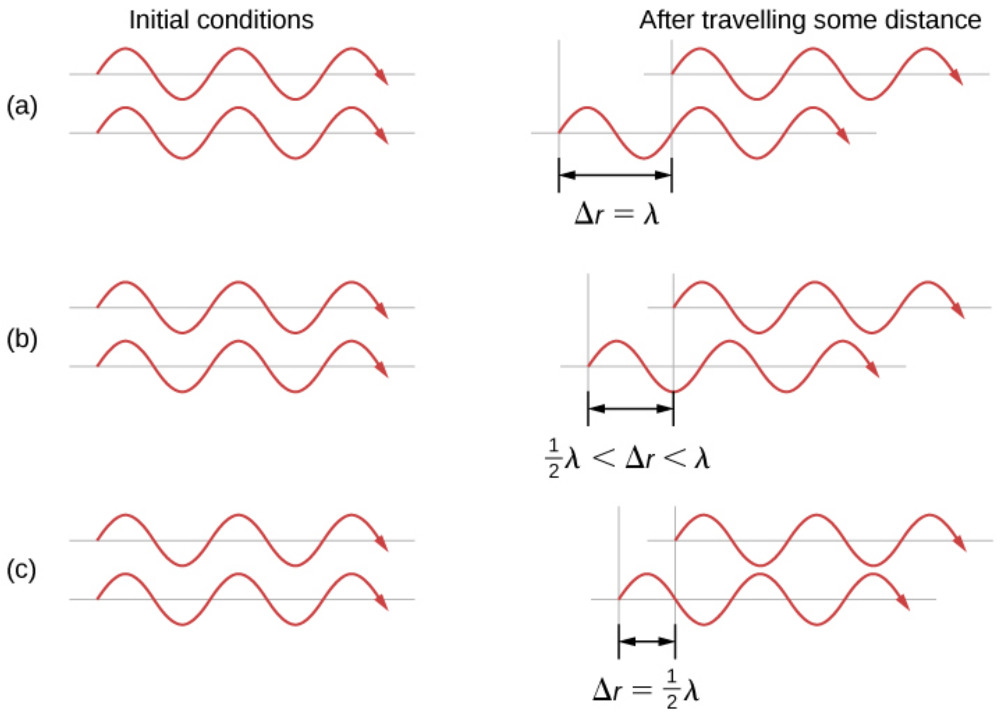
One way for two identical waves that are initially in phase to become out of phase with one another is to have the waves travel different distances; that is, they have different path lengths. Sound waves provide an excellent example of a phase shift due to a path difference. As we have discussed, sound waves can basically be modeled as longitudinal waves, where the molecules of the medium oscillate around an equilibrium position, or as pressure waves.

When the waves leave the speakers, they move out as spherical waves ([link]). The waves interfere; constructive inference is produced by the combination of two crests or two troughs, as shown. Destructive interference is produced by the combination of a trough and a crest.



When sound waves are produced by a speaker, they travel at the speed of sound and move out as spherical waves. Here, two speakers produce the same steady tone (frequency). The result is points of high-intensity sound (highlighted), which result from two crests (compression) or two troughs (rarefaction) overlapping. Destructive interference results from a crest and trough overlapping. The points where there is constructive interference in the figure occur because the two waves are in phase at those points. Points of destructive interference ([link]) are the result of the two waves being out of phase.


Crest
Trough
Speaker
P
Signal generator
Speaker
r_1
P
r_2
$\Delta r = |r_2 - r_1|$
● Purely constructive - two crests
● Purely constructive - two troughs
● Purely destructive - one crest and one trough


Initial conditions
After travelling some distance
(a)
$\Delta r = \lambda$
(b)
$\frac{1}{2}\lambda < \Delta r < \lambda$
(c)
$\Delta r = \frac{1}{2}\lambda$

Two speakers being driven by a single signal generator. The sound waves produced by the speakers are in phase and are of a single frequency. The sound waves interfere with each other. When two crests or two troughs coincide, there is constructive interference, marked by the red and blue dots. When a trough and a crest coincide, destructive interference occurs, marked by black dots. The phase difference is due to the path lengths traveled by the individual waves. Two identical waves travel two different path lengths to a point P. (a) The difference in the path lengths is one wavelength, resulting in total constructive interference and a resulting amplitude equal to twice the original amplitude. (b) The difference in the path lengths is less than one wavelength but greater than one half a wavelength, resulting in an amplitude greater than zero and less than twice the original amplitude. (c) The difference in the path lengths is one half of a wavelength, resulting in total destructive interference and a resulting amplitude of zero.

The phase difference at each point is due to the different path lengths traveled by each wave. When the difference in the path lengths is an integer multiple of a wavelength,

Equation:

$$\Delta r = |r_2 - r_1| = n\lambda,\ \text{where}\, n = 0, 1, 2, 3, \ldots,$$

the waves are in phase and there is constructive interference. When the difference in path lengths is an odd multiple of a half wavelength,

Equation:

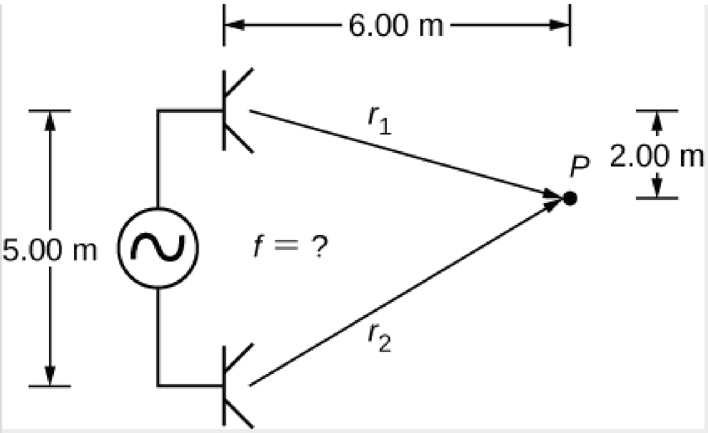
$$\Delta r = |r_2 - r_1| = n\frac{\lambda}{2},\ \text{where}\, n = 1, 3, 5, \ldots,$$

the waves are $180^\circ\,(\pi\,\text{rad})$ out of phase and the result is destructive interference. These points can be located with a sound-level intensity meter.

Example:

Interference of Sound Waves

Two speakers are separated by 5.00 m and are being driven by a signal generator at an unknown frequency. A student with a sound-level meter walks out 6.00 m and down 2.00 m, and finds the first minimum intensity, as shown below. What is the frequency supplied by the signal generator? Assume the wave speed of sound is $v = 343.00\,\text{m/s}$.

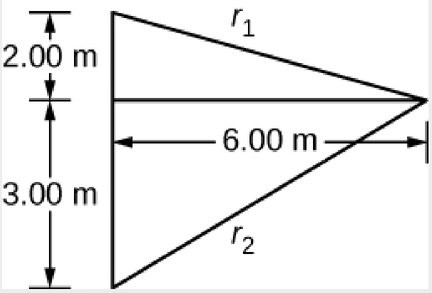


Strategy

The wave velocity is equal to $v = \frac{\lambda}{T} = \lambda f$. The frequency is then $f = \frac{v}{\lambda}$. A minimum intensity indicates destructive interference and the first such point occurs where there is path difference of $\Delta r = \lambda/2$, which can be found from the geometry.

Solution

1. Find the path length to the minimum point from each speaker.

   Equation:

   $$r_1 = \sqrt{(6.00\,\text{m})^2 + (2.00\,\text{m})^2} = 6.32\,\text{m},\ \ r_2 = \sqrt{(6.00\,\text{m})^2 + (3.00\,\text{m})^2} = 6.71\,\text{m}$$

2. Use the difference in the path length to find the wavelength.
   Equation:

   $$\Delta r = |r_2 - r_1| = |6.71\,\text{m} - 6.32\,\text{m}| = 0.39\,\text{m}$$

   Equation:

   $$\lambda = 2\Delta r = 2\,(0.39\,\text{m}) = 0.78\,\text{m}$$

3. Find the frequency.

Equation:

$$f = \frac{v}{\lambda} = \frac{343.00\ \text{m/s}}{0.78\ \text{m}} = 439.74\ \text{Hz}$$

Significance
If point P were a point of maximum intensity, then the path length would be an integer multiple of the wavelength.

Note:
Exercise:

Problem:

Check Your Understanding If you walk around two speakers playing music, how come you do not notice places where the music is very loud or very soft, that is, where there is constructive and destructive interference?

Solution:

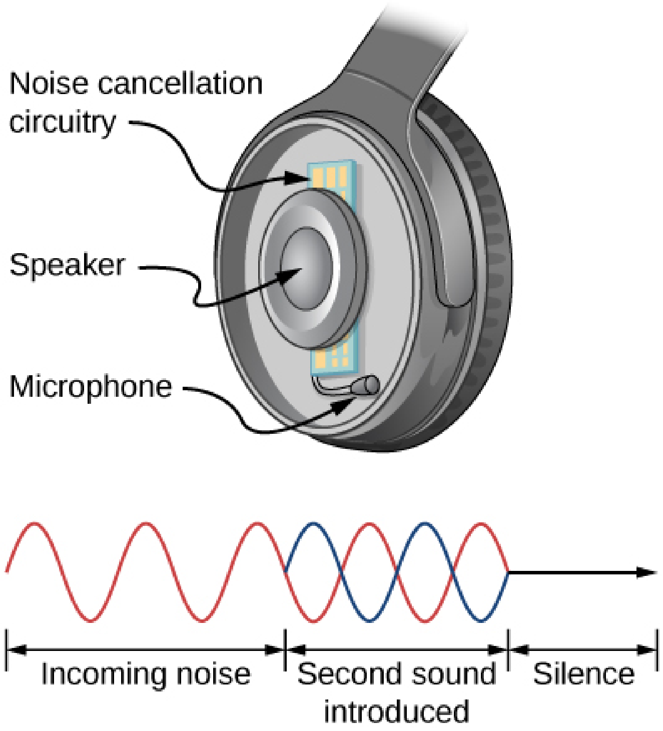
In the example, the two speakers were producing sound at a single frequency. Music has various frequencies and wavelengths.

The concept of a phase shift due to a difference in path length is very important. You will use this concept again in Interference and Photons and Matter Waves, where we discuss how Thomas Young used this method in his famous double-slit experiment to provide evidence that light has wavelike properties.

Noise Reduction through Destructive Interference

[link] shows a clever use of sound interference to cancel noise. Larger-scale applications of active noise reduction by destructive interference have been proposed for entire passenger compartments in commercial aircraft. To obtain destructive interference, a fast electronic analysis is performed, and a second sound is introduced 180° out of phase with the original sound, with its maxima and minima exactly reversed from the incoming noise. Sound waves in fluids are pressure waves and are consistent with Pascal’s principle; that is, pressures from two different sources add and subtract like simple numbers. Therefore, positive and negative gauge pressures add to a much smaller pressure, producing a lower-intensity sound. Although completely destructive interference is possible only under the

simplest conditions, it is possible to reduce noise levels by 30 dB or more using this technique.


Noise cancellation circuitry
Speaker
Microphone
Incoming noise
Second sound introduced
Silence

Headphones designed to cancel noise with destructive interference create a sound wave exactly opposite to the incoming sound. These headphones can be more effective than the simple passive attenuation used in most ear protection. Such headphones were used on the record-setting, around-the-world nonstop flight of the *Voyager* aircraft in 1986 to protect the pilots’ hearing from engine noise.

Note:
Exercise:

Problem:

Check Your Understanding Describe how noise-canceling headphones differ from standard headphones used to block outside sounds.

Solution:

Regular headphones only block sound waves with a physical barrier. Noise-canceling headphones use destructive interference to reduce the loudness of outside sounds.

Where else can we observe sound interference? All sound resonances, such as in musical instruments, are due to constructive and destructive interference. Only the resonant frequencies interfere constructively to form standing waves, whereas others interfere destructively and are absent.

Resonance in a Tube Closed at one End

As we discussed in Waves, *standing waves* are formed by two waves moving in opposite directions. When two identical sinusoidal waves move in opposite directions, the waves may be modeled as

Equation:

$$y_1\left(x,t\right)=A\sin\left(kx-\omega t\right)\text{ and }y_2\left(x,t\right)=A\sin\left(kx+\omega t\right).$$

When these two waves interfere, the resultant wave is a standing wave:

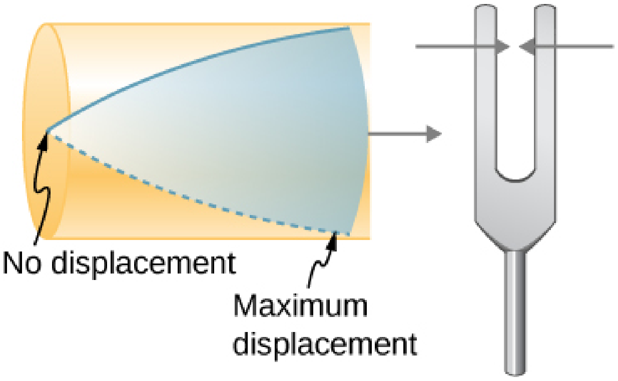
Equation:

$$y_{\text{R}}\left(x,t\right)=\left[2A\sin\left(kx\right)\right]\cos\left(\omega t\right).$$

Resonance can be produced due to the boundary conditions imposed on a wave. In Waves, we showed that resonance could be produced in a string under tension that had symmetrical boundary conditions, specifically, a node at each end. We defined a node as a fixed point where the string did not move. We found that the symmetrical boundary conditions resulted in some frequencies resonating and producing standing waves, while other frequencies interfere destructively. Sound waves can resonate in a hollow tube, and the frequencies of the sound waves that resonate depend on the boundary conditions.

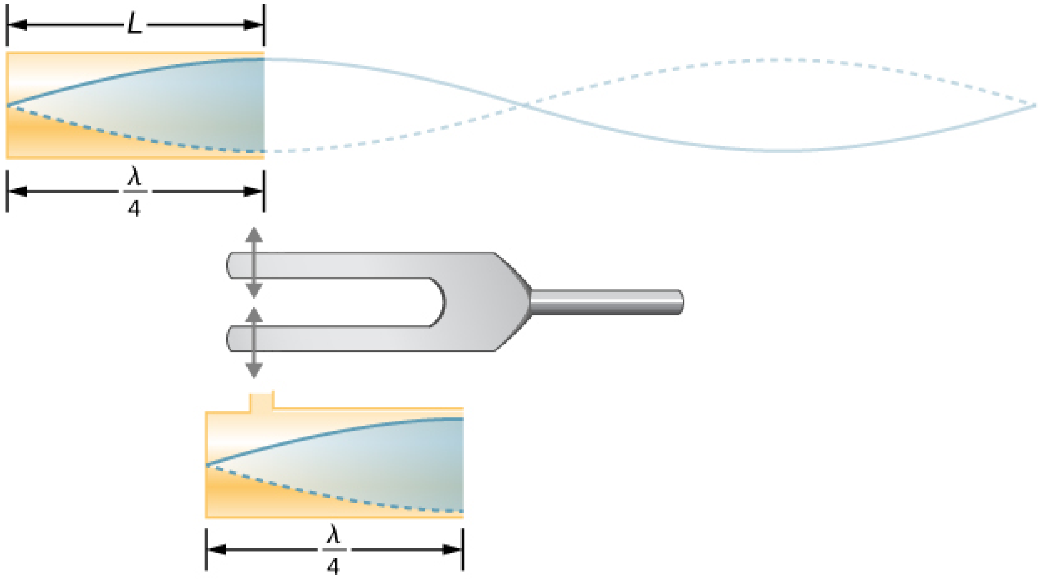
Suppose we have a tube that is closed at one end and open at the other. If we hold a vibrating tuning fork near the open end of the tube, an incident sound wave travels through the tube and reflects off the closed end. The reflected sound has the same frequency and wavelength as the incident sound wave, but is traveling in the opposite direction. At the closed end of the tube, the molecules of air have very little freedom to oscillate, and a node arises. At the open end, the molecules are free to move, and at the right frequency, an antinode occurs. Unlike the symmetrical boundary conditions for the standing waves on the string, the boundary conditions for a tube open at one end and closed at the other end are anti-symmetrical: a node at the closed end and an antinode at the open end.

If the tuning fork has just the right frequency, the air column in the tube resonates loudly, but at most frequencies it vibrates very little. This observation just means that the air column has only certain natural frequencies. Consider the lowest frequency that will cause the tube to resonate, producing a loud sound. There will be a node at the closed end and an antinode at the open end, as shown in [link].



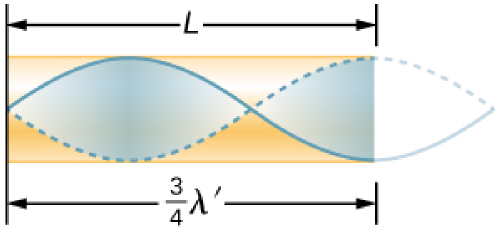
Resonance of air in a tube closed at one end, caused by a tuning fork that vibrates at the lowest frequency that can produce resonance (the fundamental frequency). A node exists at the closed end and an antinode at the open end.

The standing wave formed in the tube has an antinode at the open end and a node at the closed end. The distance from a node to an antinode is one-fourth of a wavelength, and this equals the length of the tube; thus, $\lambda_1 = 4L$. This same resonance can be produced by a vibration introduced at or near the closed end of the tube ([link]). It is best to consider this a natural vibration of the air column, independently of how it is induced.



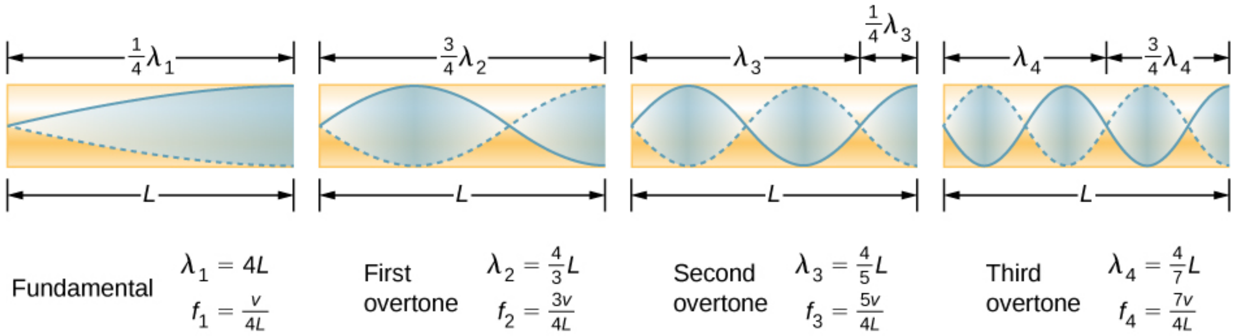
The same standing wave is created in the tube by a vibration introduced near its closed end.

Given that maximum air displacements are possible at the open end and none at the closed end, other shorter wavelengths can resonate in the tube, such as the one shown in [link]. Here the standing wave has three-fourths of its wavelength in the tube, or $\frac{3}{4}\lambda_3 = L$, so that $\lambda_3 = \frac{4}{3}L$. Continuing this process reveals a whole series of shorter-wavelength and higher-frequency sounds that resonate in the tube. We use specific terms for the resonances in any system. The lowest resonant frequency is called the **fundamental**, while all higher resonant frequencies are called **overtones**. The resonant frequencies that are integral multiples of the fundamental are collectively called **harmonics**. The fundamental is the first harmonic, the second harmonic is twice the frequency of the first harmonic, and so on. Some of these harmonics may not exist for a given scenario. [link] shows the fundamental and the first three overtones (or the first, third, fifth, and seventh harmonics) in a tube closed at one end.



Another resonance for a tube

closed at one end. This standing wave has maximum air displacement at the open end and none at the closed end. The wavelength is shorter, with three-fourths λ' equaling the length of the tube, so that $\lambda' = 4L/3$. This higher-frequency vibration is the first overtone.


$\frac{1}{4}\lambda_1$
$\frac{3}{4}\lambda_2$
λ_3
$\frac{1}{4}\lambda_3$
λ_4
$\frac{3}{4}\lambda_4$
L
L
L
L
Fundamental
$\lambda_1 = 4L$
$f_1 = \frac{v}{4L}$
First overtone
$\lambda_2 = \frac{4}{3}L$
$f_2 = \frac{3v}{4L}$
Second overtone
$\lambda_3 = \frac{4}{5}L$
$f_3 = \frac{5v}{4L}$
Third overtone
$\lambda_4 = \frac{4}{7}L$
$f_4 = \frac{7v}{4L}$

The fundamental and three lowest overtones for a tube closed at one end. All have maximum air displacements at the open end and none at the closed end.

The relationship for the resonant wavelengths of a tube closed at one end is

Note:
Equation:

$$\lambda_n = \frac{4}{n}L \quad n = 1, 3, 5, \ldots$$

Now let us look for a pattern in the resonant frequencies for a simple tube that is closed at one end. The fundamental has $\lambda = 4L$, and frequency is related to wavelength and the speed of sound as given by

$v = f\lambda.$

Solving for *f* in this equation gives

Equation:

$$f = \frac{v}{\lambda} = \frac{v}{4L},$$

where *v* is the speed of sound in air. Similarly, the first overtone has $\lambda = 4L/3$ (see [link]), so that

Equation:

$$f_3 = 3\frac{v}{4L} = 3f_1.$$

Because $f_3 = 3f_1$, we call the first overtone the third harmonic. Continuing this process, we see a pattern that can be generalized in a single expression. The resonant frequencies of a tube closed at one end are

Note:

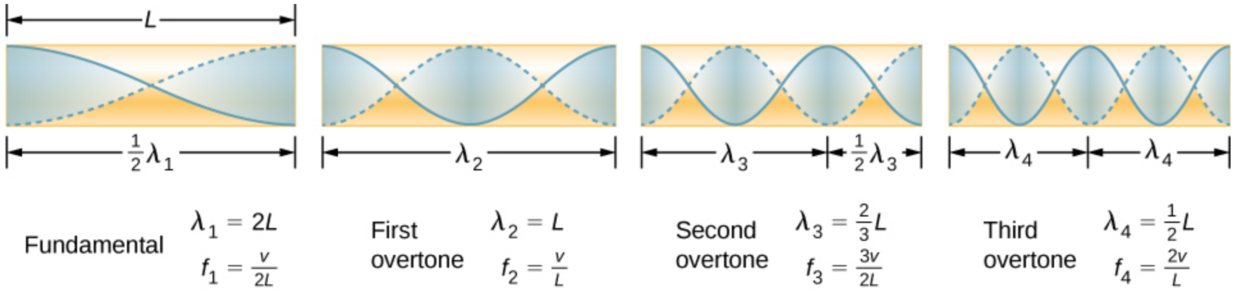
Equation:

$$f_n = n\frac{v}{4L}, \quad n = 1, 3, 5, ...,$$

where f_1 is the fundamental, f_3 is the first overtone, and so on. It is interesting that the resonant frequencies depend on the speed of sound and, hence, on temperature. This dependence poses a noticeable problem for organs in old unheated cathedrals, and it is also the reason why musicians commonly bring their wind instruments to room temperature before playing them.

Resonance in a Tube Open at Both Ends

Another source of standing waves is a tube that is open at both ends. In this case, the boundary conditions are symmetrical: an antinode at each end. The resonances of tubes open at both ends can be analyzed in a very similar fashion to those for tubes closed at one end. The air columns in tubes open at both ends have maximum air displacements at both ends ([link]). Standing waves form as shown.



L
$\frac{1}{2}\lambda_1$
λ_2
λ_3
$\frac{1}{2}\lambda_3$
λ_4
λ_4
Fundamental
$\lambda_1 = 2L$
$f_1 = \frac{v}{2L}$
First overtone
$\lambda_2 = L$
$f_2 = \frac{v}{L}$
Second overtone
$\lambda_3 = \frac{2}{3}L$
$f_3 = \frac{3v}{2L}$
Third overtone
$\lambda_4 = \frac{1}{2}L$
$f_4 = \frac{2v}{L}$

The resonant frequencies of a tube open at both ends, including the fundamental and the first three overtones. In all cases, the maximum air displacements occur at both ends of the tube, giving it different natural frequencies than a tube closed at one end.

The relationship for the resonant wavelengths of a tube open at both ends is

Note:
Equation:

$$\lambda_n = \frac{2}{n}L, \quad n = 1, 2, 3,....$$

Based on the fact that a tube open at both ends has maximum air displacements at both ends, and using [link] as a guide, we can see that the resonant frequencies of a tube open at both ends are

Note:
Equation:

$$f_n = n\frac{v}{2L}, \quad n = 1, 2, 3...,$$

where f_1 is the fundamental, f_2 is the first overtone, f_3 is the second overtone, and so on. Note that a tube open at both ends has a fundamental frequency twice what it would have if closed at one end. It also has a different spectrum of overtones than a tube closed at one end.

Note that a tube open at both ends has symmetrical boundary conditions, similar to the string fixed at both ends discussed in Waves. The relationships for the wavelengths and frequencies of a stringed instrument are the same as given in [link] and [link]. The speed of the wave on the string (from Waves) is $v=\sqrt{\frac{F_T}{\mu}}$. The air around the string vibrates at the same frequency as the string, producing sound of the same frequency. The sound wave moves at the speed of sound and the wavelength can be found using $v=\lambda f$.

Note:
Exercise:

Problem:

Check Your Understanding How is it possible to use a standing wave's node and antinode to determine the length of a closed-end tube?

Solution:

When the tube resonates at its natural frequency, the wave's node is located at the closed end of the tube, and the antinode is located at the open end. The length of the tube is equal to one-fourth of the wavelength of this wave. Thus, if we know the wavelength of the wave, we can determine the length of the tube.

Note:
This video lets you visualize sound waves.

Note:
Exercise:

Problem:

Check Your Understanding You observe two musical instruments that you cannot identify. One plays high-pitched sounds and the other plays low-pitched sounds. How could you determine which is which without hearing either of them play?

Solution:

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Summary

- Unwanted sound can be reduced using destructive interference.
- Sound has the same properties of interference and resonance as defined for all waves.
- In air columns, the lowest-frequency resonance is called the fundamental, whereas all higher resonant frequencies are called overtones. Collectively, they are called harmonics.

Conceptual Questions

Exercise:

Problem:

You are given two wind instruments of identical length. One is open at both ends, whereas the other is closed at one end. Which is able to produce the lowest frequency?

Solution:

The fundamental wavelength of a tube open at each end is $2L$, where the wavelength of a tube open at one end and closed at one end is $4L$. The tube open at one end has the lower fundamental frequency, assuming the speed of sound is the same in both tubes.

Exercise:

Problem:

What is the difference between an overtone and a harmonic? Are all harmonics overtones? Are all overtones harmonics?

Exercise:

Problem:

Two identical columns, open at both ends, are in separate rooms. In room *A*, the temperature is $T = 20^\circ\text{C}$ and in room *B*, the temperature is $T = 25^\circ\text{C}$. A speaker is attached to the end of each tube, causing the tubes to resonate at the fundamental frequency. Is the frequency the same for both tubes? Which has the higher frequency?

Solution:

The wavelength in each is twice the length of the tube. The frequency depends on the wavelength and the speed of the sound waves. The frequency in room *B* is higher because the speed of sound is higher where the temperature is higher.

Problems

Exercise:

Problem:

(a) What is the fundamental frequency of a 0.672-m-long tube, open at both ends, on a day when the speed of sound is 344 m/s? (b) What is the frequency of its second harmonic?

Exercise:

Problem:

What is the length of a tube that has a fundamental frequency of 176 Hz and a first overtone of 352 Hz if the speed of sound is 343 m/s?

Solution:

0.974 m

Exercise:

Problem:

The ear canal resonates like a tube closed at one end. (See [link]Figure 17_03_HumEar[/link].) If ear canals range in length from 1.80 to 2.60 cm in an average population, what is the range of fundamental resonant frequencies? Take air temperature to be 37.0°C, which is the same as body temperature.

Exercise:

Problem:

Calculate the first overtone in an ear canal, which resonates like a 2.40-cm-long tube closed at one end, by taking air temperature to be 37.0°C. Is the ear particularly sensitive to such a frequency? (The resonances of the ear canal are complicated by its nonuniform shape, which we shall ignore.)

Solution:

11.0 kHz; The ear is not particularly sensitive to this frequency, so we don’t hear overtones due to the ear canal.

Exercise:

Problem:

A crude approximation of voice production is to consider the breathing passages and mouth to be a resonating tube closed at one end. (a) What is the fundamental frequency if the tube is 0.240 m long, by taking air temperature to be 37.0°C? (b) What would this frequency become if the person replaced the air with helium? Assume the same temperature dependence for helium as for air.

Exercise:

Problem:

A 4.0-m-long pipe, open at one end and closed at one end, is in a room where the temperature is $T = 22^\circ\text{C}$. A speaker capable of producing variable frequencies is placed at the open end and is used to cause the tube to resonate. (a) What is the wavelength and the frequency of the fundamental frequency? (b) What is the frequency and wavelength of the first overtone?

Solution:

a. $v = 344.08\,\text{m/s},\ \ \lambda_1 = 16.00\,\text{m},\ \ f_1 = 21.51\,\text{Hz}$;
b. $\lambda_3 = 5.33\,\text{m},\ \ f_3 = 64.56\,\text{Hz}$

Exercise:

Problem:

A 4.0-m-long pipe, open at both ends, is placed in a room where the temperature is $T = 25^\circ\text{C}$. A speaker capable of producing variable frequencies is placed at the open end and is used to cause the tube to resonate. (a) What are the wavelength and the frequency of the fundamental frequency? (b) What are the frequency and wavelength of the first overtone?

Exercise:

Problem:

A nylon guitar string is fixed between two lab posts 2.00 m apart. The string has a linear mass density of $\mu = 7.20\,\text{g/m}$ and is placed under a tension of 160.00 N. The string is placed next to a tube, open at both ends, of length L. The string is plucked and the tube resonates at the $n = 3$ mode. The speed of sound is 343 m/s. What is the length of the tube?

Solution:

$$v_{\text{string}} = 149.07\,\text{m/s},\ \ \lambda_3 = 1.33\,\text{m},\ \ f_3 = 112.08\,\text{Hz}$$
$$\lambda_1 = \frac{v}{f_1},\ \ L = 1.53\,\text{m}$$

Exercise:

Problem:

A 512-Hz tuning fork is struck and placed next to a tube with a movable piston, creating a tube with a variable length. The piston is slid down the pipe and resonance is reached when the piston is 115.50 cm from the open end. The next resonance is reached when the piston is 82.50 cm from the open end. (a) What is the speed of sound in the tube? (b) How far from the open end will the piston cause the next mode of resonance?

Exercise:

Problem:

Students in a physics lab are asked to find the length of an air column in a tube closed at one end that has a fundamental frequency of 256 Hz. They hold the tube vertically and fill it with water to the top, then lower the water while a 256-Hz tuning fork is rung and listen for the first resonance. (a) What is the air temperature if the resonance occurs for a length of 0.336 m? (b) At what length will they observe the second resonance (first overtone)?

Solution:

a. $22.0\text{°}\text{C}$; b. 1.01 m

Glossary

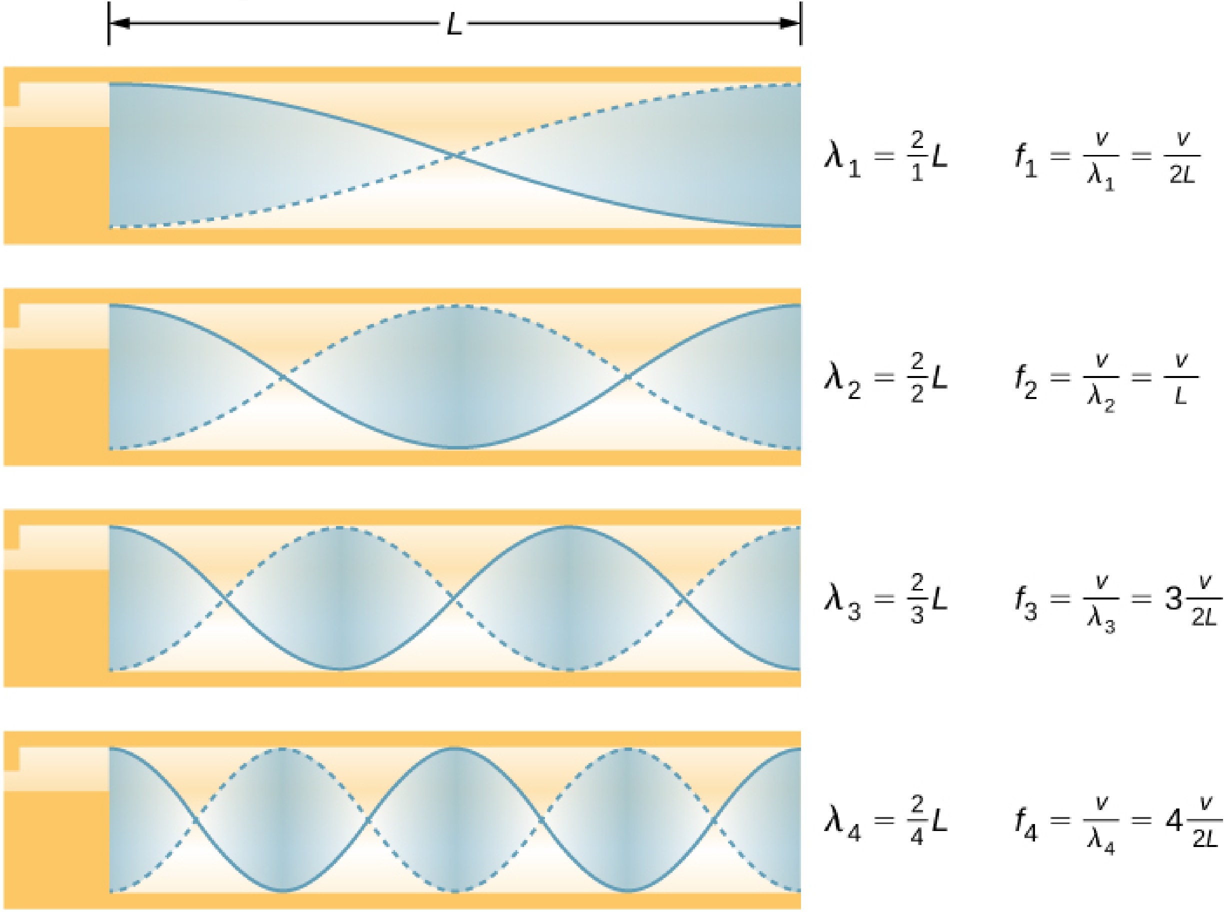
fundamental	the lowest-frequency resonance
harmonics	the term used to refer collectively to the fundamental and its overtones
overtones	all resonant frequencies higher than the fundamental

Sources of Musical Sound

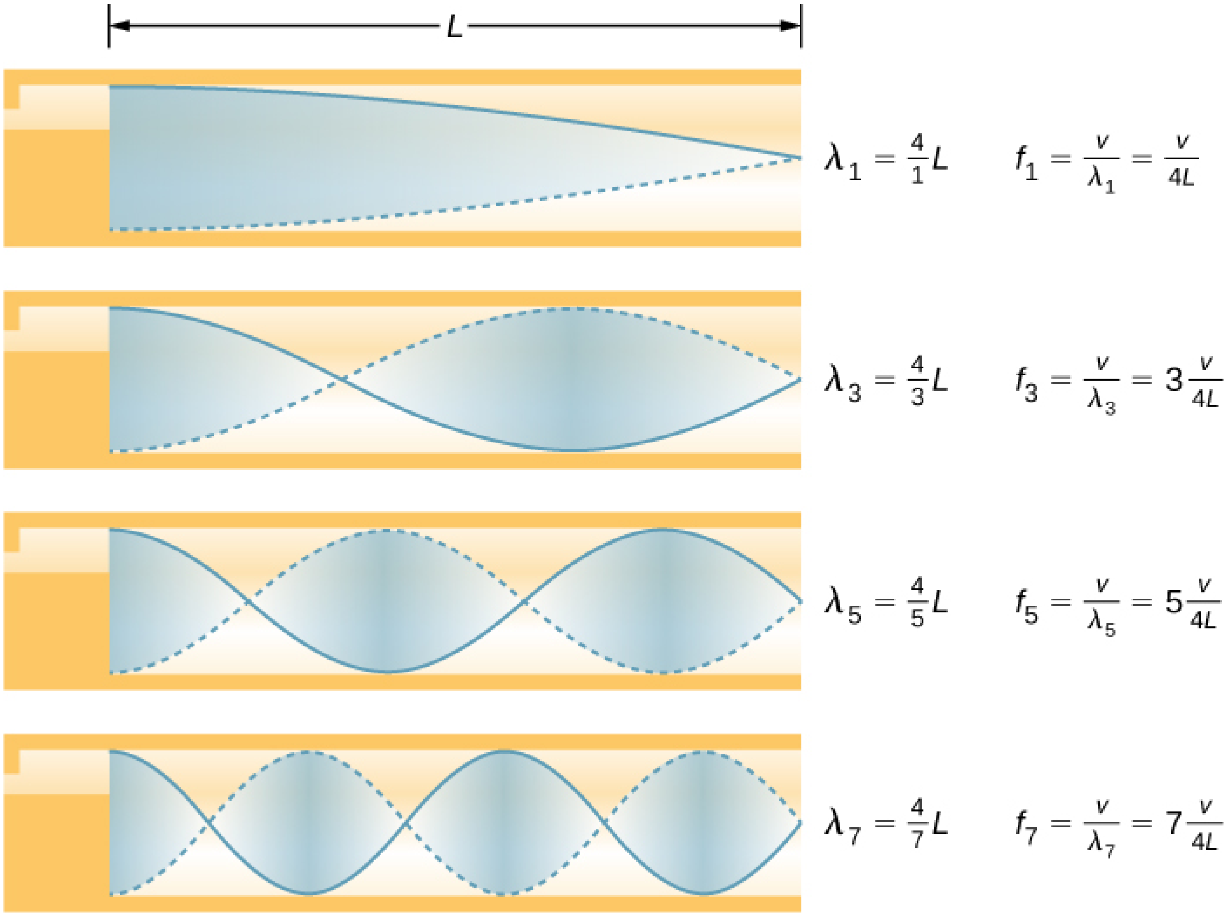
By the end of this section, you will be able to:

- Describe the resonant frequencies in instruments that can be modeled as a tube with symmetrical boundary conditions
- Describe the resonant frequencies in instruments that can be modeled as a tube with anti-symmetrical boundary conditions

Some musical instruments, such as woodwinds, brass, and pipe organs, can be modeled as tubes with symmetrical boundary conditions, that is, either open at both ends or closed at both ends ([link]). Other instruments can be modeled as tubes with anti-symmetrical boundary conditions, such as a tube with one end open and the other end closed ([link]).


L
$\lambda_1 = \frac{2}{1}L$
$f_1 = \frac{v}{\lambda_1} = \frac{v}{2L}$
$\lambda_2 = \frac{2}{2}L$
$f_2 = \frac{v}{\lambda_2} = \frac{v}{L}$
$\lambda_3 = \frac{2}{3}L$
$f_3 = \frac{v}{\lambda_3} = 3\frac{v}{2L}$
$\lambda_4 = \frac{2}{4}L$
$f_4 = \frac{v}{\lambda_4} = 4\frac{v}{2L}$

Some musical instruments can be modeled as a pipe open at both ends.



L
$\lambda_1 = \frac{4}{1}L$
$f_1 = \frac{v}{\lambda_1} = \frac{v}{4L}$
$\lambda_3 = \frac{4}{3}L$
$f_3 = \frac{v}{\lambda_3} = 3\frac{v}{4L}$
$\lambda_5 = \frac{4}{5}L$
$f_5 = \frac{v}{\lambda_5} = 5\frac{v}{4L}$
$\lambda_7 = \frac{4}{7}L$
$f_7 = \frac{v}{\lambda_7} = 7\frac{v}{4L}$

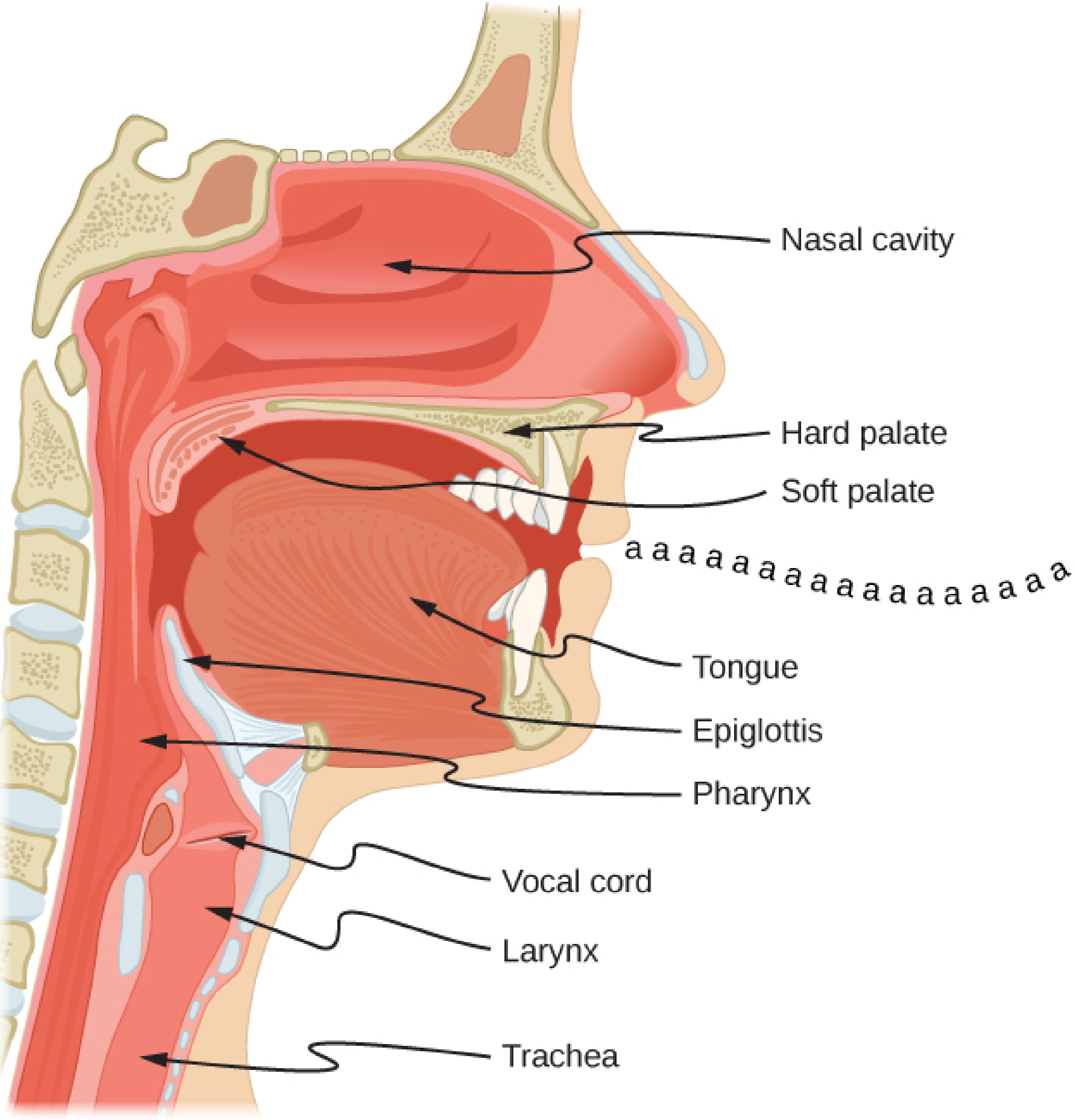
Some musical instruments can be modeled as a pipe closed at one end.

Resonant frequencies are produced by longitudinal waves that travel down the tubes and interfere with the reflected waves traveling in the opposite direction. A pipe organ is manufactured with various tubes of fixed lengths to produce different frequencies. The waves are the result of compressed air allowed to expand in the tubes. Even in open tubes, some reflection occurs due to the constraints of the sides of the tubes and the atmospheric pressure outside the open tube.

The antinodes do not occur at the opening of the tube, but rather depend on the radius of the tube. The waves do not fully expand until they are outside the open end of a tube, and for a thin-walled tube, an *end correction* should be added. This end correction is approximately 0.6 times the radius of the tube and should be added to the length of the tube.

Players of instruments such as the flute or oboe vary the length of the tube by opening and closing finger holes. On a trombone, you change the tube length by using a sliding tube. Bugles have a fixed length and can produce only a limited range of frequencies.

The fundamental and overtones can be present simultaneously in a variety of combinations. For example, middle C on a trumpet sounds distinctively different from middle C on a clarinet, although both instruments are modified versions of a tube closed at one end. The fundamental frequency is the same (and usually the most intense), but the overtones and their mix of intensities are different and subject to shading by the musician. This mix is what gives various musical instruments (and human voices) their distinctive characteristics, whether they have air columns, strings, sounding boxes, or drumheads. In fact, much of our speech is determined by shaping the cavity formed by the throat and mouth, and positioning the tongue to adjust the fundamental and combination of overtones. For example, simple resonant cavities can be made to resonate with the sound of the vowels ([link]). In boys at puberty, the larynx grows and the shape of the resonant cavity changes, giving rise to the difference in predominant frequencies in speech between men and women.


Nasal cavity
Hard palate
Soft palate
a a a a a a a a a a a a a a a a a
Tongue
Epiglottis
Pharynx
Vocal cord
Larynx
Trachea

The throat and mouth form an air column closed at one end that resonates in response to vibrations in the voice box. The spectrum of overtones and their intensities vary with mouth shaping and tongue position to form different sounds. The voice box can be replaced with a mechanical vibrator, and understandable speech is still possible. Variations in basic shapes make different voices recognizable.

Example:

Finding the Length of a Tube with a 128-Hz Fundamental

(a) What length should a tube closed at one end have on a day when the air temperature is
22.0°C if its fundamental frequency is to be 128 Hz (C below middle C)?
(b) What is the frequency of its fourth overtone?

Strategy

The length *L* can be found from the relationship $f_n = n\frac{v}{4L}$, but we first need to find the speed of sound *v*.

Solution

a. Identify knowns: The fundamental frequency is 128 Hz, and the air temperature is 22.0°C.
   Use $f_n = n\frac{v}{4L}$ to find the fundamental frequency ($n = 1$),
   Equation:

$$f_1 = \frac{v}{4L}.$$

   Solve this equation for length,
   Equation:

$$L = \frac{v}{4f_1}.$$

   Find the speed of sound using $v = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}$,
   Equation:

$$v = (331\ \text{m/s})\sqrt{\frac{295\ \text{K}}{273\ \text{K}}} = 344\ \text{m/s}.$$

   Enter the values of the speed of sound and frequency into the expression for *L*.
   Equation:

$$L = \frac{v}{4f_1} = \frac{344\,\text{m/s}}{4(128\,\text{Hz})} = 0.672\,\text{m}$$

b. Identify knowns: The first overtone has $n = 3$, the second overtone has $n = 5$, the third overtone has $n = 7$, and the fourth overtone has $n = 9$.
Enter the value for the fourth overtone into $f_n = n\frac{v}{4L}$,
Equation:

$$f_9 = 9\frac{v}{4L} = 9f_1 = 1.15\,\text{kHz}.$$

Significance
Many wind instruments are modified tubes that have finger holes, valves, and other devices for changing the length of the resonating air column and hence, the frequency of the note played. Horns producing very low frequencies require tubes so long that they are coiled into loops. An example is the tuba. Whether an overtone occurs in a simple tube or a musical instrument depends on how it is stimulated to vibrate and the details of its shape. The trombone, for example, does not produce its fundamental frequency and only makes overtones.

If you have two tubes with the same fundamental frequency, but one is open at both ends and the other is closed at one end, they would sound different when played because they have different overtones. Middle C, for example, would sound richer played on an open tube, because it has even multiples of the fundamental as well as odd. A closed tube has only odd multiples.

Resonance

Resonance occurs in many different systems, including strings, air columns, and atoms. As we discussed in earlier chapters, resonance is the driven or forced oscillation of a system at its natural frequency. At resonance, energy is transferred rapidly to the oscillating system, and the amplitude of its oscillations grows until the system can no longer be described by Hooke’s law. An example of this is the distorted sound intentionally produced in certain types of rock music.

Wind instruments use resonance in air columns to amplify tones made by lips or vibrating reeds. Other instruments also use air resonance in clever ways to amplify sound. [link] shows a violin and a guitar, both of which have sounding boxes but with different shapes, resulting in different overtone structures. The vibrating string creates a sound that resonates in the sounding box, greatly amplifying the sound and creating overtones that give the instrument its characteristic timbre. The more complex the shape of the sounding box, the greater its ability to resonate over a wide range of frequencies. The marimba, like the one shown in [link], uses pots or gourds below the wooden slats to amplify their tones. The resonance of the pot can be adjusted by adding water.

String instruments such as (a) violins and (b) guitars use resonance in their sounding boxes to amplify and enrich the sound created by their vibrating strings. The bridge and supports couple the string vibrations to the sounding boxes and air within. (credit a: modification of work by Feliciano Guimarães; credit b: modification of work by Steve Snodgrass)

Resonance has been used in musical instruments since prehistoric times. This marimba uses gourds as resonance chambers to amplify its sound. (credit: “APC Events”/Flickr)

We have emphasized sound applications in our discussions of resonance and standing waves, but these ideas apply to any system that has wave characteristics. Vibrating strings, for example, are actually resonating and have fundamentals and overtones similar to those for air columns. More subtle are the resonances in atoms due to the wave character of their electrons. Their orbitals can be viewed as standing waves, which have a fundamental (ground state) and overtones (excited states). It is fascinating that wave characteristics apply to such a wide range of physical systems.

Summary

- Some musical instruments can be modeled as pipes that have symmetrical boundary conditions: open at both ends or closed at both ends. Other musical instruments can be modeled as pipes that have anti-symmetrical boundary conditions: closed at one end and open at the other.
- Some instruments, such as the pipe organ, have several tubes with different lengths. Instruments such as the flute vary the length of the tube by closing the holes along the tube. The trombone varies the length of the tube using a sliding bar.
- String instruments produce sound using a vibrating string with nodes at each end. The air around the string oscillates at the frequency of the string. The relationship for the frequencies for the string is the same as for the symmetrical boundary conditions of the pipe, with the length of the pipe replaced by the length of the string and the velocity replaced by $v = \sqrt{\frac{F_T}{\mu}}$.

Conceptual Questions

Exercise:

Problem:

How does an unamplified guitar produce sounds so much more intense than those of a plucked string held taut by a simple stick?

Exercise:

Problem:

Consider three pipes of the same length (L). Pipe A is open at both ends, pipe B is closed at both ends, and pipe C has one open end and one closed end. If the velocity of sound is the same in each of the three tubes, in which of the tubes could the lowest fundamental frequency be produced? In which of the tubes could the highest fundamental frequency be produced?

Solution:

When resonating at the fundamental frequency, the wavelength for pipe *C* is $4L$, and for pipes *A* and *B* is $2L$. The frequency is equal to $f = v/\lambda$. Pipe *C* has the lowest frequency and pipes *A* and *B* have equal frequencies, higher than the one in pipe *C*.

Exercise:

Problem:

Pipe *A* has a length *L* and is open at both ends. Pipe *B* has a length *L*/2 and has one open end and one closed end. Assume the speed of sound to be the same in both tubes. Which of the harmonics in each tube would be equal?

Exercise:

Problem:

A string is tied between two lab posts a distance *L* apart. The tension in the string and the linear mass density is such that the speed of a wave on the string is $v = 343\,\text{m/s}$. A tube with symmetric boundary conditions has a length *L* and the speed of sound in the tube is $v = 343\,\text{m/s}$. What could be said about the frequencies of the harmonics in the string and the tube? What if the velocity in the string were $v = 686\,\text{m/s}$?

Solution:

Since the boundary conditions are both symmetric, the frequencies are $f_n = \frac{nv}{2L}$. Since the speed is the same in each, the frequencies are the same. If the wave speed were doubled in the string, the frequencies in the string would be twice the frequencies in the tube.

Problems

Exercise:

Problem:

If a wind instrument, such as a tuba, has a fundamental frequency of 32.0 Hz, what are its first three overtones? It is closed at one end. (The overtones of a real tuba are more complex than this example, because it is a tapered tube.)

Exercise:

Problem:

What are the first three overtones of a bassoon that has a fundamental frequency of 90.0 Hz? It is open at both ends. (The overtones of a real bassoon are more complex than this example, because its double reed makes it act more like a tube closed at one end.)

Solution:

$$\text{first overtone} = 180\text{ Hz};$$
$$\text{second overtone} = 270\text{ Hz};$$
$$\text{third overtone} = 360\text{ Hz}$$

Exercise:

Problem:

How long must a flute be in order to have a fundamental frequency of 262 Hz (this frequency corresponds to middle C on the evenly tempered chromatic scale) on a day when air temperature is 20.0°C? It is open at both ends.

Exercise:

Problem:

What length should an oboe have to produce a fundamental frequency of 110 Hz on a day when the speed of sound is 343 m/s? It is open at both ends.

Solution:

1.56 m

Exercise:

Problem:

(a) Find the length of an organ pipe closed at one end that produces a fundamental frequency of 256 Hz when air temperature is $18.0\text{°}\text{C}$. (b) What is its fundamental frequency at $25.0\text{°}\text{C}$?

Exercise:

Problem:

An organ pipe $(L = 3.00\,\text{m})$ is closed at both ends. Compute the wavelengths and frequencies of the first three modes of resonance. Assume the speed of sound is$v = 343.00\,\text{m/s}$.

Solution:

The pipe has symmetrical boundary conditions;

$$\lambda_n = \frac{2}{n}L, \quad f_n = \frac{nv}{2L}, \quad n = 1, 2, 3$$
$$\lambda_1 = 6.00\,\text{m}, \quad \lambda_2 = 3.00\,\text{m}, \quad \lambda_3 = 2.00\,\text{m}$$
$$f_1 = 57.17\,\text{Hz}, \quad f_2 = 114.33\,\text{Hz}, \quad f_3 = 171.50\,\text{Hz}$$

Exercise:

Problem:

An organ pipe $(L = 3.00\,\text{m})$ is closed at one end. Compute the wavelengths and frequencies of the first three modes of resonance. Assume the speed of sound is$v = 343.00\,\text{m/s}$.

Exercise:

Problem:

A sound wave of a frequency of 2.00 kHz is produced by a string oscillating in the $n = 6$ mode. The linear mass density of the string is $\mu = 0.0065\,\text{kg/m}$ and the length of the string is 1.50 m. What is the tension in the string?

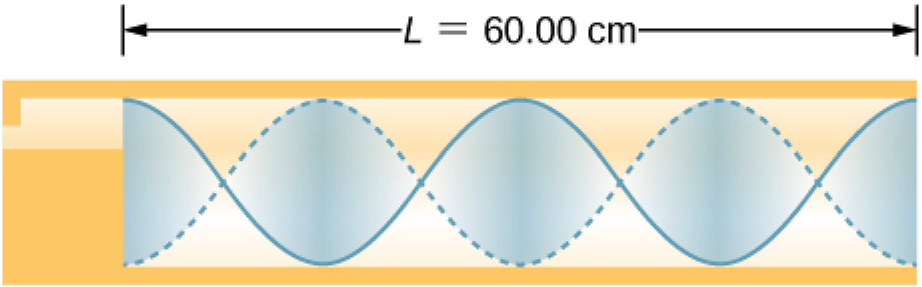
Solution:

$$\lambda_6 = 0.5\,\text{m}$$
$$v = 1000\,\text{m/s}$$
$$F_T = 6500\,\text{N}$$

Exercise:

Problem:

Consider the sound created by resonating the tube shown below. The air temperature is $T_\text{C} = 30.00^\circ\text{C}$. What are the wavelength, wave speed, and frequency of the sound produced?



Exercise:

Problem:

A student holds an 80.00-cm lab pole one quarter of the length from the end of the pole. The lab pole is made of aluminum. The student strikes the lab pole with a hammer. The pole resonates at the lowest possible frequency. What is that frequency?

Solution:

$f = 6.40\,\text{kHz}$

Exercise:

Problem:

A string on the violin has a length of 24.00 cm and a mass of 0.860 g. The fundamental frequency of the string is 1.00 kHz. (a) What is the speed of the wave on the string? (b) What is the tension in the string?

Exercise:

Problem:

By what fraction will the frequencies produced by a wind instrument change when air temperature goes from 10.0º C to 30.0º C? That is, find the ratio of the frequencies at those temperatures.

Solution:

1.03 or 3%

Beats

By the end of this section, you will be able to:

- Determine the beat frequency produced by two sound waves that differ in frequency
- Describe how beats are produced by musical instruments

The study of music provides many examples of the superposition of waves and the constructive and destructive interference that occurs. Very few examples of music being performed consist of a single source playing a single frequency for an extended period of time. You will probably agree that a single frequency of sound for an extended period might be boring to the point of irritation, similar to the unwanted drone of an aircraft engine or a loud fan. Music is pleasant and interesting due to mixing the changing frequencies of various instruments and voices.

An interesting phenomenon that occurs due to the constructive and destructive interference of two or more frequencies of sound is the phenomenon of **beats**. If two sounds differ in frequencies, the sound waves can be modeled as

Equation:

$$y_1 = A\cos\left(k_1x - 2\pi f_1t\right) \text{ and } y_2 = A\cos\left(k_2x - 2\pi f_2t\right).$$

Using the trigonometric identity $\cos u + \cos v = 2\cos\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right)$ and considering the point in space as $x = 0.0\,\text{m}$, we find the resulting sound at a point in space, from the superposition of the two sound waves, is equal to [link]:

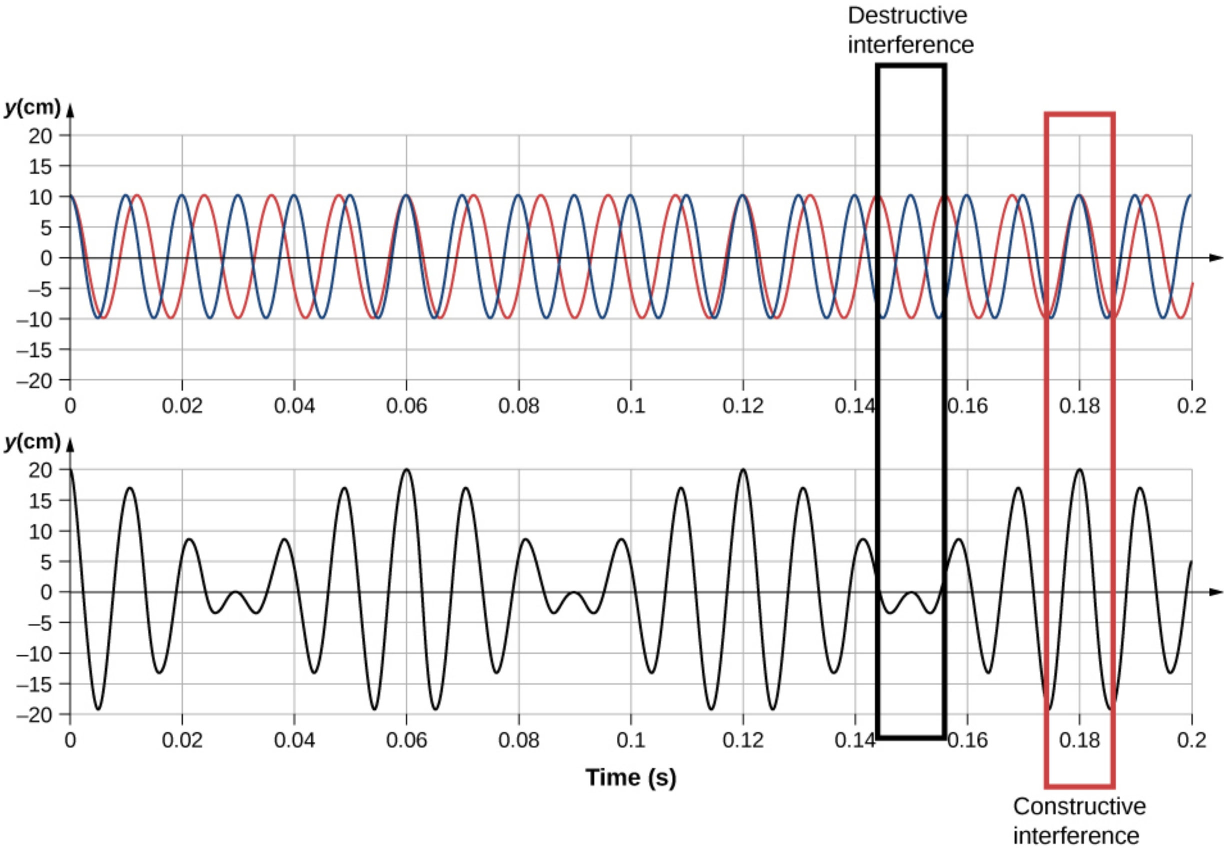
Equation:

$$y\left(t\right) = 2A\cos\left(2\pi f_{avg}t\right)\cos\left(2\pi\left(\frac{|f_2 - f_1|}{2}\right)t\right),$$

where the **beat frequency** is

Note:
Equation:

$$f_{\text{beat}} = |f_2 - f_1|.$$



Beats produced by the constructive and destructive interference of two sound waves that differ in frequency.

These beats can be used by piano tuners to tune a piano. A tuning fork is struck and a note is played on the piano. As the piano tuner tunes the string, the beats have a lower frequency as the frequency of the note played approaches the frequency of the tuning fork.

Example:

Find the Beat Frequency Between Two Tuning Forks

What is the beat frequency produced when a tuning fork of a frequency of 256 Hz and a tuning fork of a frequency of 512 Hz are struck simultaneously?

Strategy

The beat frequency is the difference of the two frequencies.

Solution

We use $f_{\text{beat}} = |f_2 - f_1|$:

Equation:

$$|f_2 - f_1| = (512 - 256)\,\text{Hz} = 256\,\text{Hz}.$$

Significance

The beat frequency is the absolute value of the difference between the two frequencies. A negative frequency would not make sense.

Note:

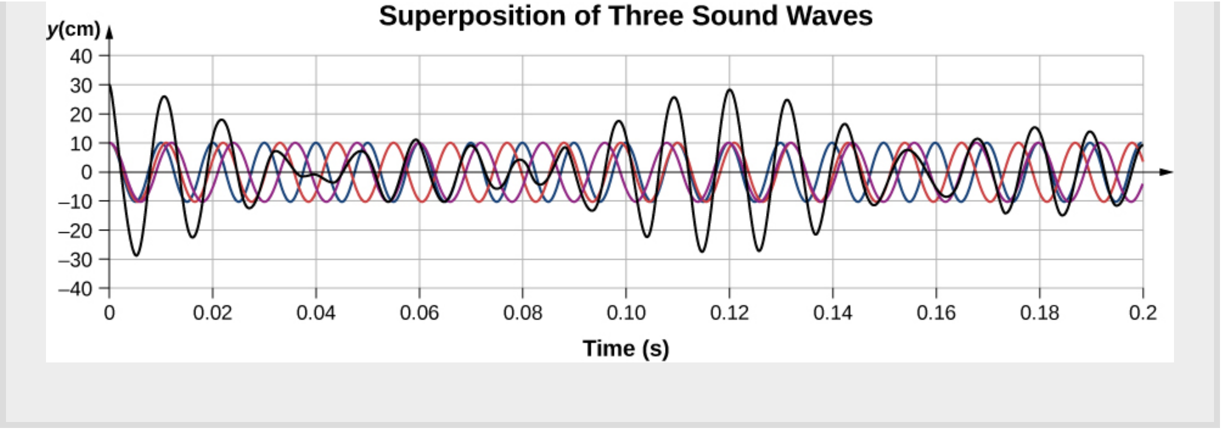
Exercise:

Problem:

Check Your Understanding What would happen if more than two frequencies interacted? Consider three frequencies.

Solution:

An easy way to understand this event is to use a graph, as shown below. It appears that beats are produced, but with a more complex pattern of interference.



The study of the superposition of various waves has many interesting applications beyond the study of sound. In later chapters, we will discuss the wave properties of particles. The particles can be modeled as a “wave packet” that results from the superposition of various waves, where the particle moves at the “group velocity” of the wave packet.

Summary

- When two sound waves that differ in frequency interfere, beats are created with a beat frequency that is equal to the absolute value of the difference in the frequencies.

Conceptual Questions

Exercise:

Problem:

Two speakers are attached to variable-frequency signal generator. Speaker *A* produces a constant-frequency sound wave of 1.00 kHz, and speaker *B* produces a tone of 1.10 kHz. The beat frequency is 0.10 kHz. If the frequency of each speaker is doubled, what is the beat frequency produced?

Exercise:

Problem:

The label has been scratched off a tuning fork and you need to know its frequency. From its size, you suspect that it is somewhere around 250 Hz. You find a 250-Hz tuning fork and a 270-Hz tuning fork. When you strike the 250-Hz fork and the fork of unknown frequency, a beat frequency of 5 Hz is produced. When you strike the unknown with the 270-Hz fork, the beat frequency is 15 Hz. What is the unknown frequency? Could you have deduced the frequency using just the 250-Hz fork?

Solution:

The frequency of the unknown fork is 255 Hz. No, if only the 250 Hz fork is used, listening to the beat frequency could only limit the possible frequencies to 245 Hz or 255 Hz.

Exercise:

Problem:

Referring to the preceding question, if you had only the 250-Hz fork, could you come up with a solution to the problem of finding the unknown frequency?

Exercise:

Problem:

A “showy” custom-built car has two brass horns that are supposed to produce the same frequency but actually emit 263.8 and 264.5 Hz. What beat frequency is produced?

Solution:

The beat frequency is 0.7 Hz.

Problems

Exercise:

Problem:

What beat frequencies are present: (a) If the musical notes A and C are played together (frequencies of 220 and 264 Hz)? (b) If D and F are played together (frequencies of 297 and 352 Hz)? (c) If all four are played together?

Exercise:

Problem:

What beat frequencies result if a piano hammer hits three strings that emit frequencies of 127.8, 128.1, and 128.3 Hz?

Solution:

$$\begin{aligned} f_{\text{B}} &= |f_1 - f_2| \\ |128.3\ \text{Hz} - 128.1\ \text{Hz}| &= 0.2\ \text{Hz}; \\ |128.3\ \text{Hz} - 127.8\ \text{Hz}| &= 0.5\ \text{Hz}; \\ |128.1\ \text{Hz} - 127.8\ \text{Hz}| &= 0.3\ \text{Hz} \end{aligned}$$

Exercise:

Problem:

A piano tuner hears a beat every 2.00 s when listening to a 264.0-Hz tuning fork and a single piano string. What are the two possible frequencies of the string?

Exercise:

Problem:

Two identical strings, of identical lengths of 2.00 m and linear mass density of $\mu = 0.0065\ \text{kg/m}$, are fixed on both ends. String A is under a tension of 120.00 N. String B is under a tension of 130.00 N. They are each plucked and produce sound at the $n = 10$ mode. What is the beat frequency?

Solution:

$$v_A = 135.87\,\mathrm{m/s},\ v_B = 141.42\,\mathrm{m/s},$$
$$\lambda_A = \lambda_B = 0.40\,\mathrm{m}$$
$$\Delta f = 15.00\,\mathrm{Hz}$$

Exercise:

Problem:

A piano tuner uses a 512-Hz tuning fork to tune a piano. He strikes the fork and hits a key on the piano and hears a beat frequency of 5 Hz. He tightens the string of the piano, and repeats the procedure. Once again he hears a beat frequency of 5 Hz. What happened?

Exercise:

Problem:

A string with a linear mass density of $\mu = 0.0062\,\mathrm{kg/m}$ is stretched between two posts 1.30 m apart. The tension in the string is 150.00 N. The string oscillates and produces a sound wave. A 1024-Hz tuning fork is struck and the beat frequency between the two sources is 52.83 Hz. What are the possible frequency and wavelength of the wave on the string?

Solution:

$$v = 155.54\,\mathrm{m/s},$$
$$f_{\text{string}} = 971.17\,\mathrm{Hz},\ n = 16.23$$
$$f_{\text{string}} = 1076.83\,\mathrm{Hz},\ n = 18.00$$

The frequency is 1076.83 Hz and the wavelength is 0.14 m.

Exercise:

Problem:

A car has two horns, one emitting a frequency of 199 Hz and the other emitting a frequency of 203 Hz. What beat frequency do they produce?

Exercise:

Problem:

The middle C hammer of a piano hits two strings, producing beats of 1.50 Hz. One of the strings is tuned to 260.00 Hz. What frequencies could the other string have?

Solution:

$$f_2 = f_1 \pm f_{\text{B}} = 260.00\ \text{Hz} \pm 1.50\ \text{Hz},$$
$$\text{so that } f_2 = 261.50\ \text{Hz or } f_2 = 258.50\ \text{Hz}$$

Exercise:

Problem:

Two tuning forks having frequencies of 460 and 464 Hz are struck simultaneously. What average frequency will you hear, and what will the beat frequency be?

Exercise:

Problem:

Twin jet engines on an airplane are producing an average sound frequency of 4100 Hz with a beat frequency of 0.500 Hz. What are their individual frequencies?

Solution:

$f_{\text{ace}} = \frac{f_1 + f_2}{2}; f_{\text{B}} = f_1 - f_2 (\text{assume} f_1 > f_2)$

$f_{\text{ace}} = \frac{(f_{\text{B}} + f_2) + f_2}{2} \Rightarrow$

$f_2 = 4099.750\,\text{Hz}$

$f_1 = 4100.250\,\text{Hz}$

Exercise:

Problem:

Three adjacent keys on a piano (F, F-sharp, and G) are struck simultaneously, producing frequencies of 349, 370, and 392 Hz. What beat frequencies are produced by this discordant combination?

Glossary

beat frequency
frequency of beats produced by sound waves that differ in frequency

beats
constructive and destructive interference of two or more frequencies of sound

The Doppler Effect

By the end of this section, you will be able to:

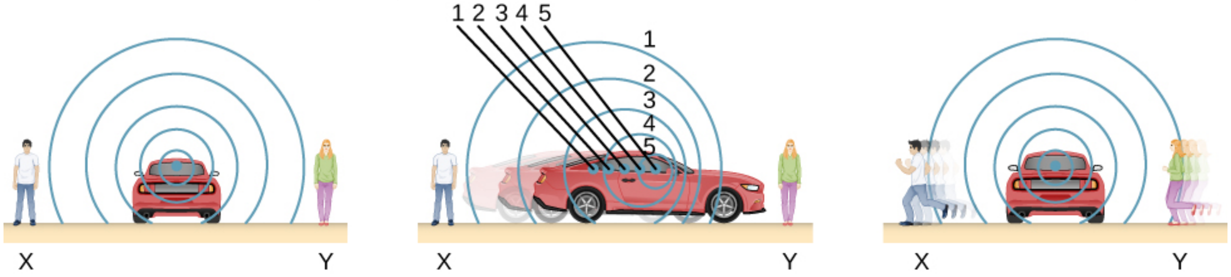
- Explain the change in observed frequency as a moving source of sound approaches or departs from a stationary observer
- Explain the change in observed frequency as an observer moves toward or away from a stationary source of sound

The characteristic sound of a motorcycle buzzing by is an example of the **Doppler effect**. Specifically, if you are standing on a street corner and observe an ambulance with a siren sounding passing at a constant speed, you notice two characteristic changes in the sound of the siren. First, the sound increases in loudness as the ambulance approaches and decreases in loudness as it moves away, which is expected. But in addition, the high-pitched siren shifts dramatically to a lower-pitched sound. As the ambulance passes, the frequency of the sound heard by a stationary observer changes from a constant high frequency to a constant lower frequency, even though the siren is producing a constant source frequency. The closer the ambulance brushes by, the more abrupt the shift. Also, the faster the ambulance moves, the greater the shift. We also hear this characteristic shift in frequency for passing cars, airplanes, and trains.

The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning horn, you will hear the horn’s frequency shift from high to low as you pass by. The actual change in frequency due to relative motion of source and observer is called a **Doppler shift**. The Doppler effect and Doppler shift are named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? [link] illustrates sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point at which the sound is emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave are centered on the same point, and the stationary observers on either side hear the same wavelength and frequency as emitted by the source (case a). If the source is moving, the situation is different. Each compression of the air moves out in a sphere from the point at which it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other.

Thus, the wavelength is shorter in the direction the source is moving (on the right in case b), and longer in the opposite direction (on the left in case b). Finally, if the observers move, as in case (c), the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.



(a) Stationary source and observers

(b) Moving source, stationary observers

(c) Stationary source, moving observers

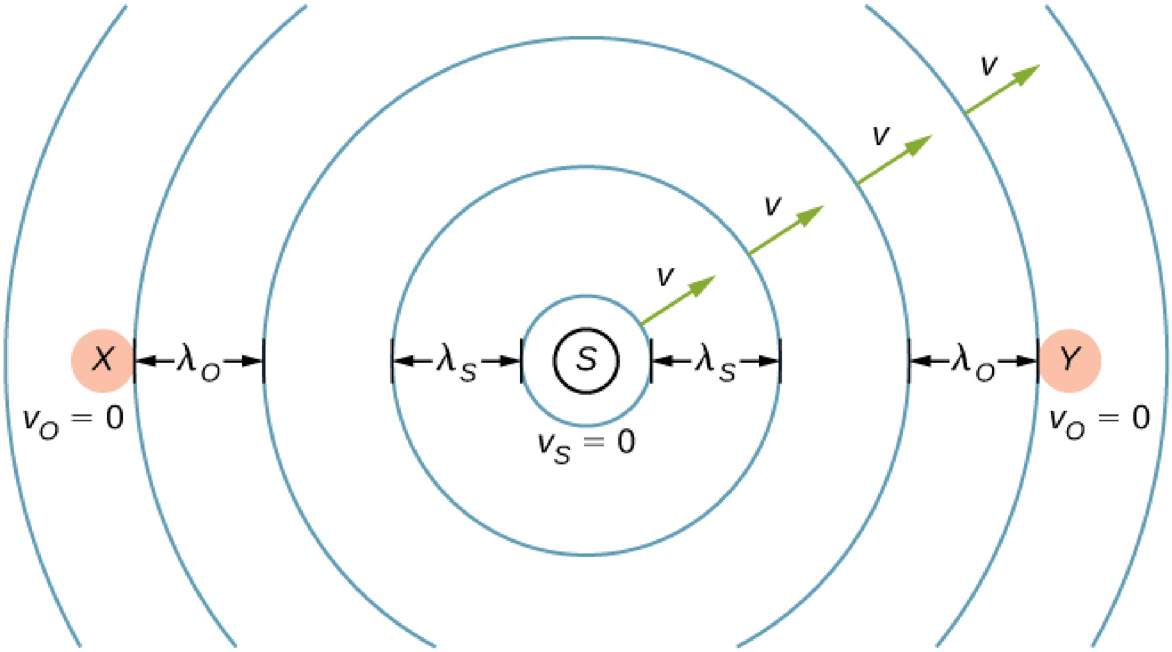
Sounds emitted by a source spread out in spherical waves. (a) When the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers. (b) Sounds emitted by a source moving to the right spread out from the points at which they were emitted. The wavelength is reduced, and consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitched sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced. (c) The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v = f\lambda$, where v is the fixed speed of sound. The sound moves in a medium and has the same speed v in that medium whether the source is moving or not. Thus, f multiplied by λ is a constant. Because the observer on the right in case (b) receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in case (c). A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed, the greater the effect.

The Doppler effect occurs not only for sound, but for any wave when there is relative motion between the observer and the source. Doppler shifts occur in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The relative velocities of stars and galaxies is determined by the shift in the frequencies of light received from them and has implied much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

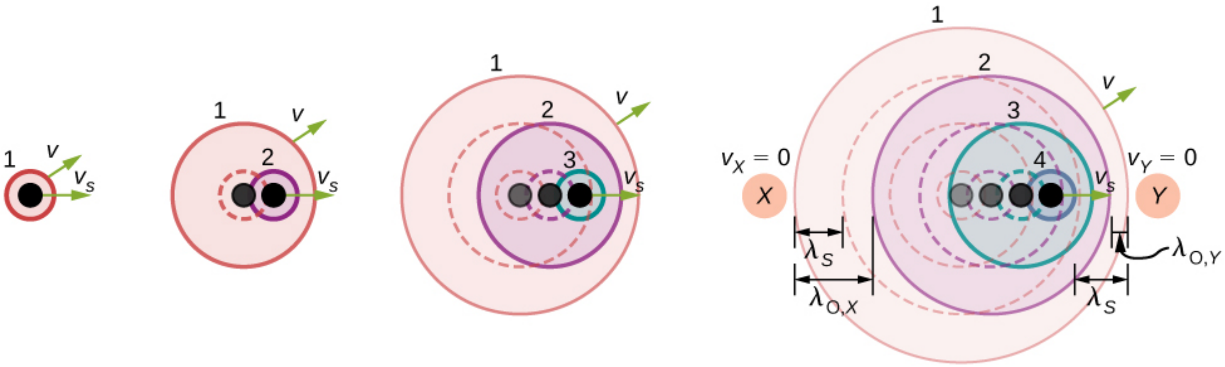
Derivation of the Observed Frequency due to the Doppler Shift

Consider two stationary observers *X* and *Y* in [link], located on either side of a stationary source. Each observer hears the same frequency, and that frequency is the frequency produced by the stationary source.



A stationary source sends out sound waves at a constant frequency f_s, with a constant wavelength λ_s, at the speed of sound v. Two stationary observers X and Y, on either side of the source, observe a frequency $f_o = f_s$, with a wavelength $\lambda_o = \lambda_s$.

Now consider a stationary observer X with a source moving away from the observer with a constant speed $v_s < v$ ([link]). At time $t = 0$, the source sends out a sound wave, indicated in black. This wave moves out at the speed of sound *v*. The position of the sound wave at each time interval of period T_s is shown as dotted lines. After one period, the source has moved $\Delta x = v_s T_s$ and emits a second sound wave, which moves out at the speed of sound. The source continues to move and produce sound waves, as indicated by the circles numbered 3 and 4. Notice that as the waves move out, they remained centered at their respective point of origin.



A source moving at a constant speed v_s away from an observer X. The moving source sends out sound waves at a constant frequency f_s, with a constant wavelength λ_s, at the speed of sound *v*. Snapshots of the source at an interval of T_s are shown as the source moves away from the stationary observer X. The solid lines represent the position of the sound waves after four periods from the initial time. The dotted lines are used to show the positions of the waves at each time period. The observer hears a wavelength of $\lambda_o = \lambda_s + \Delta x = \lambda_s + v_s T_s$.

Using the fact that the wavelength is equal to the speed times the period, and the period is the inverse of the frequency, we can derive the observed frequency:

Equation:

$$\begin{aligned}
\lambda_o &= \lambda_s + \Delta x \\
vT_o &= vT_s + v_s T_s \\
\frac{v}{f_o} &= \frac{v}{f_s} = \frac{v_s}{f_s} = \frac{v+v_s}{f_s} \\
f_o &= f_s\left(\frac{v}{v+v_s}\right).
\end{aligned}$$

As the source moves away from the observer, the observed frequency is lower than the source frequency.

Now consider a source moving at a constant velocity v_s, moving toward a stationary observer *Y*, also shown in [link]. The wavelength is observed by *Y* as $\lambda_\text{o} = \lambda_\text{s} - \Delta x = \lambda_\text{s} - v_\text{s} T_\text{s}$. Once again, using the fact that the wavelength is equal to the speed times the period, and the period is the inverse of the frequency, we can derive the observed frequency:

Equation:

$$\begin{array}{rcl} \lambda_\text{o} & = & \lambda_\text{s} - \Delta x \\ vT_\text{o} & = & vT_\text{s} - v_\text{s} T_\text{s} \\ \frac{v}{f_\text{o}} & = & \frac{v}{f_\text{s}} - \frac{v_\text{s}}{f_\text{s}} = \frac{v - v_\text{s}}{f_\text{s}} \\ f_\text{o} & = & f_\text{s} \left(\frac{v}{v - v_\text{s}} \right). \end{array}$$

When a source is moving and the observer is stationary, the observed frequency is

Note:
Equation:

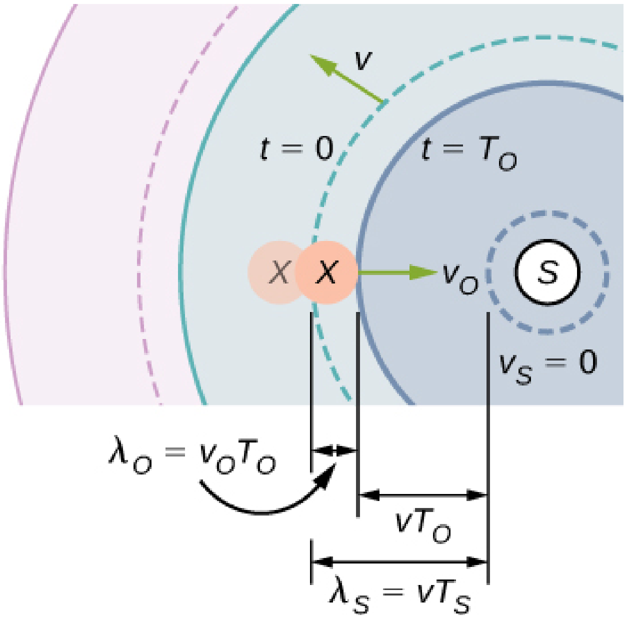
$$f_\text{o} = f_\text{s} \left(\frac{v}{v \mp v_\text{s}} \right),$$

where f_o is the frequency observed by the stationary observer, f_s is the frequency produced by the moving source, *v* is the speed of sound, v_s is the constant speed of the source, and the top sign is for the source approaching the observer and the bottom sign is for the source departing from the observer.

What happens if the observer is moving and the source is stationary? If the observer moves toward the stationary source, the observed frequency is higher than the source frequency. If the observer is moving away from the stationary source, the observed frequency is lower than the source frequency. Consider observer *X* in [link] as the observer moves toward a stationary source with a speed v_o. The source emits a tone with a constant frequency f_s and constant period T_s. The observer hears the first wave emitted by the source. If the observer were stationary, the time for one wavelength of sound to pass should be equal to the period of the source T_s. Since the observer is moving toward the source, the time for one wavelength to pass is less than T_s and is equal to the observed period $T_o = T_s - \Delta t$. At time $t = 0$, the observer starts at the beginning of a wavelength and moves toward the second wavelength as the wavelength moves out from the source. The wavelength is equal to the distance the observer traveled plus the distance the sound wave traveled until it is met by the observer:

Equation:

$$\begin{aligned}
\lambda_s &= vT_o + v_oT_o \\
vT_s &= (v + v_o)T_o \\
v\left(\frac{1}{f_s}\right) &= (v + v_o)\left(\frac{1}{f_o}\right) \\
f_o &= f_s\left(\frac{v + v_o}{v}\right).
\end{aligned}$$

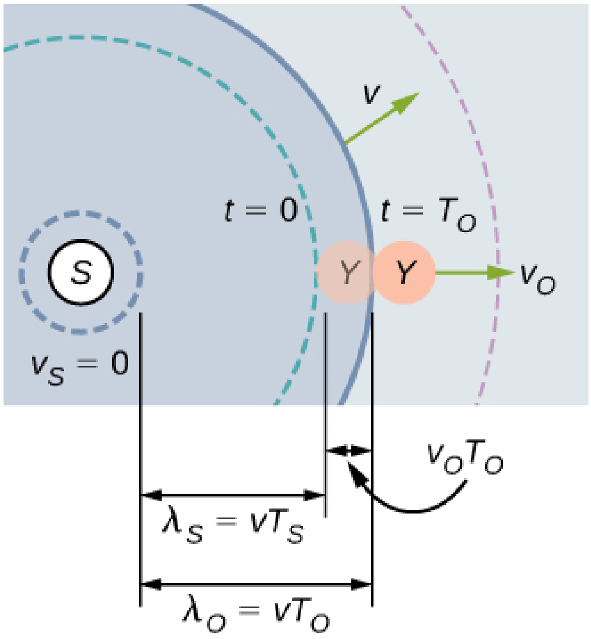


A stationary source emits a sound wave with a constant frequency f_s, with a constant wavelength λ_s moving at the speed of sound v. Observer X moves toward the source with a constant speed v_o, and the figure shows the initial and final position of observer X. Observer X observes a frequency higher than the source frequency. The solid lines show the position of the waves at $t=0$. The dotted lines show the position of the waves at $t=T_\text{o}$.

If the observer is moving away from the source ([link]), the observed frequency can be found:

Equation:

$$\begin{aligned}
\lambda_\text{s} &= vT_\text{o} - v_\text{o}T_\text{o} \\
vT_\text{s} &= (v - v_\text{o})T_\text{o} \\
v\left(\frac{1}{f_\text{s}}\right) &= (v - v_\text{o})\left(\frac{1}{f_\text{o}}\right) \\
f_\text{o} &= f_\text{s}\left(\frac{v-v_\text{o}}{v}\right).
\end{aligned}$$



A stationary source emits a sound wave with a constant frequency f_s, with a constant wavelength λ_s moving at the speed of sound v.

moving at the speed of sound v. Observer Y moves away from the source with a constant speed v_o, and the figure shows initial and final position of the observer Y. Observer Y observes a frequency lower than the source frequency. The solid lines show the position of the waves at $t = 0$. The dotted lines show the position of the waves at $t = T_o$.

The equations for an observer moving toward or away from a stationary source can be combined into one equation:

Note:
Equation:

$$f_o = f_s \left(\frac{v \pm v_o}{v} \right),$$

where f_o is the observed frequency, f_s is the source frequency, v_w is the speed of sound, v_o is the speed of the observer, the top sign is for the observer approaching the source and the bottom sign is for the observer departing from the source.

[link] and [link] can be summarized in one equation (the top sign is for approaching) and is further illustrated in [link]:

Note:
Equation:

$$f_o = f_s \left(\frac{v \pm v_o}{v \mp v_s} \right),$$

Doppler shift $f_\text{o} = f_\text{s}\left(\frac{v \pm v_\text{o}}{v \mp v_\text{s}}\right)$	**Stationary observer**	**Observer moving towards source**	**Observer moving away from source**
Stationary source	$f_\text{o} = f_\text{s}$	$f_\text{o} = f_\text{s}\left(\frac{v+v_\text{o}}{v}\right)$	$f_\text{o} = f_\text{s}\left(\frac{v-v_\text{o}}{v}\right)$
Source moving towards observer	$f_\text{o} = f_\text{s}\left(\frac{v}{v-v_\text{s}}\right)$	$f_\text{o} = f_\text{s}\left(\frac{v+v_\text{o}}{v-v_\text{s}}\right)$	$f_\text{o} = f_\text{s}\left(\frac{v-v_\text{o}}{v-v_\text{s}}\right)$
Source moving away from observer	$f_\text{o} = f_\text{s}\left(\frac{v}{v+v_\text{s}}\right)$	$f_\text{o} = f_\text{s}\left(\frac{v+v_\text{o}}{v+v_\text{s}}\right)$	$f_\text{o} = f_\text{s}\left(\frac{v-v_\text{o}}{v+v_\text{s}}\right)$

where f_o is the observed frequency, f_s is the source frequency, v_w is the speed of sound, v_o is the speed of the observer, v_s is the speed of the source, the top sign is for approaching and the bottom sign is for departing.

Note:
The Doppler effect involves motion and a video will help visualize the effects of a moving observer or source. This video shows a moving source and a stationary observer, and a moving observer and a stationary source. It also discusses the Doppler effect and its application to light.

Example:
Calculating a Doppler Shift
Suppose a train that has a 150-Hz horn is moving at 35.0 m/s in still air on a day when the speed of sound is 340 m/s.
(a) What frequencies are observed by a stationary person at the side of the tracks as the train approaches and after it passes?
(b) What frequency is observed by the train’s engineer traveling on the train?
Strategy
To find the observed frequency in (a), we must use $f_\text{obs} = f_\text{s}\left(\frac{v}{v \mp v_\text{s}}\right)$ because the source is moving. The minus sign is used for the approaching train, and the plus sign

for the receding train. In (b), there are two Doppler shifts—one for a moving source and the other for a moving observer.

Solution

a. Enter known values into $f_{\text{o}} = f_{\text{s}}\left(\frac{v}{v - v_{\text{s}}}\right)$:

   Equation:

$$f_{\text{o}} = f_{\text{s}}\left(\frac{v}{v - v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} - 35.0\text{ m/s}}\right).$$

   Calculate the frequency observed by a stationary person as the train approaches:

   Equation:

$$f_{\text{o}} = (150\text{ Hz})\,(1.11) = 167\text{ Hz}.$$

   Use the same equation with the plus sign to find the frequency heard by a stationary person as the train recedes:

   Equation:

$$f_{\text{o}} = f_{\text{s}}\left(\frac{v}{v + v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} + 35.0\text{ m/s}}\right).$$

   Calculate the second frequency:

   Equation:

$$f_{\text{o}} = (150\text{ Hz})\,(0.907) = 136\text{ Hz}.$$

b. Identify knowns:

   - It seems reasonable that the engineer would receive the same frequency as emitted by the horn, because the relative velocity between them is zero.
   - Relative to the medium (air), the speeds are $v_{\text{s}} = v_{\text{o}} = 35.0\text{ m/s}$.
   - The first Doppler shift is for the moving observer; the second is for the moving source.

   Use the following equation:

   Equation:

$$f_o = \left[f_s \left(\frac{v \pm v_o}{v} \right) \right] \left(\frac{v}{v \mp v_s} \right).$$

The quantity in the square brackets is the Doppler-shifted frequency due to a moving observer. The factor on the right is the effect of the moving source. Because the train engineer is moving in the direction toward the horn, we must use the plus sign for v_{obs}; however, because the horn is also moving in the direction away from the engineer, we also use the plus sign for v_s. But the train is carrying both the engineer and the horn at the same velocity, so $v_s = v_o$. As a result, everything but f_s cancels, yielding

Equation:

$$f_o = f_s.$$

Significance

For the case where the source and the observer are not moving together, the numbers calculated are valid when the source (in this case, the train) is far enough away that the motion is nearly along the line joining source and observer. In both cases, the shift is significant and easily noticed. Note that the shift is 17.0 Hz for motion toward and 14.0 Hz for motion away. The shifts are not symmetric.

For the engineer riding in the train, we may expect that there is no change in frequency because the source and observer move together. This matches your experience. For example, there is no Doppler shift in the frequency of conversations between driver and passenger on a motorcycle. People talking when a wind moves the air between them also observe no Doppler shift in their conversation. The crucial point is that source and observer are not moving relative to each other.

Note:

Exercise:

Problem:

Check Your Understanding Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Solution:

> If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through.

The Doppler effect and the Doppler shift have many important applications in science and engineering. For example, the Doppler shift in ultrasound can be used to measure blood velocity, and police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such “Doppler Radar” can give the velocity and direction of rain or snow in weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Summary

- The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer.
- The actual change in frequency is called the Doppler shift.

Conceptual Questions

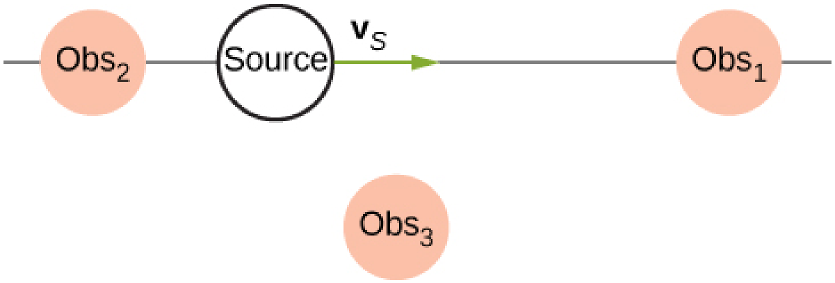
Exercise:

Problem: Is the Doppler shift real or just a sensory illusion?

Exercise:

Problem:

Three stationary observers observe the Doppler shift from a source moving at a constant velocity. The observers are stationed as shown below. Which observer will observe the highest frequency? Which observer will observe the lowest frequency? What can be said about the frequency observed by observer 3?



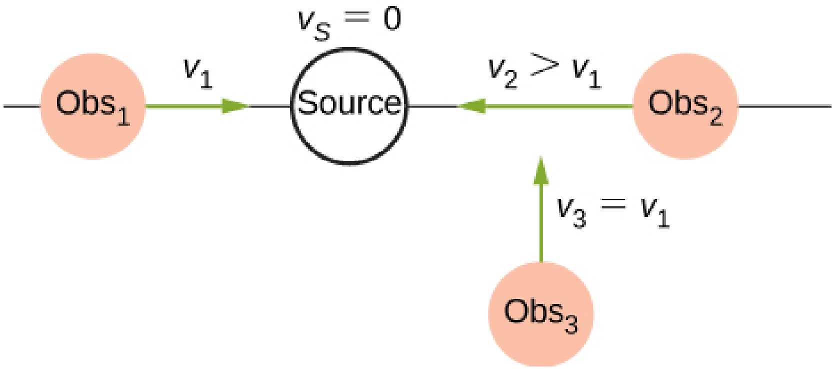
Solution:

Observer 1 will observe the highest frequency. Observer 2 will observe the lowest frequency. Observer 3 will hear a higher frequency than the source frequency, but lower than the frequency observed by observer 1, as the source approaches and a lower frequency than the source frequency, but higher than the frequency observed by observer 1, as the source moves away from observer 3.

Exercise:

Problem:

Shown below is a stationary source and moving observers. Describe the frequencies observed by the observers for this configuration.



Exercise:

Problem:

Prior to 1980, conventional radar was used by weather forecasters. In the 1960s, weather forecasters began to experiment with Doppler radar. What do you think is the advantage of using Doppler radar?

Solution:

Doppler radar can not only detect the distance to a storm, but also the speed and direction at which the storm is traveling.

Problems

Exercise:

Problem:

(a) What frequency is received by a person watching an oncoming ambulance moving at 110 km/h and emitting a steady 800-Hz sound from its siren? The speed of sound on this day is 345 m/s. (b) What frequency does she receive after the ambulance has passed?

Solution:

a. 878 Hz; b. 735 Hz

Exercise:

Problem:

(a) At an air show a jet flies directly toward the stands at a speed of 1200 km/h, emitting a frequency of 3500 Hz, on a day when the speed of sound is 342 m/s. What frequency is received by the observers? (b) What frequency do they receive as the plane flies directly away from them?

Exercise:

Problem:

What frequency is received by a mouse just before being dispatched by a hawk flying at it at 25.0 m/s and emitting a screech of frequency 3500 Hz? Take the speed of sound to be 331 m/s.

Solution:

$3.79\ \times\ 10^3\ \text{Hz}$

Exercise:

Problem:

A spectator at a parade receives an 888-Hz tone from an oncoming trumpeter who is playing an 880-Hz note. At what speed is the musician approaching if the speed of sound is 338 m/s?

Exercise:

Problem:

A commuter train blows its 200-Hz horn as it approaches a crossing. The speed of sound is 335 m/s. (a) An observer waiting at the crossing receives a frequency of 208 Hz. What is the speed of the train? (b) What frequency does the observer receive as the train moves away?

Solution:

a. 12.9 m/s; b. 193 Hz

Exercise:

Problem:

Can you perceive the shift in frequency produced when you pull a tuning fork toward you at 10.0 m/s on a day when the speed of sound is 344 m/s? To answer this question, calculate the factor by which the frequency shifts and see if it is greater than 0.300%.

Exercise:

Problem:

Two eagles fly directly toward one another, the first at 15.0 m/s and the second at 20.0 m/s. Both screech, the first one emitting a frequency of 3200 Hz and the second one emitting a frequency of 3800 Hz. What frequencies do they receive if the speed of sound is 330 m/s?

Solution:

The first eagle hears 4.23×10^3 Hz. The second eagle hears 3.56×10^3 Hz.

Exercise:

Problem:

Student *A* runs down the hallway of the school at a speed of $v_\text{o} = 5.00 \, \text{m/s}$, carrying a ringing 1024.00-Hz tuning fork toward a concrete wall. The speed of sound is $v = 343.00 \, \text{m/s}$. Student *B* stands at rest at the wall. (a) What is the frequency heard by student *B*? (b) What is the beat frequency heard by student *A*?

Exercise:

Problem:

An ambulance with a siren $(f = 1.00\text{kHz})$ blaring is approaching an accident scene. The ambulance is moving at 70.00 mph. A nurse is approaching the scene from the opposite direction, running at $v_o = 7.00\,\text{m/s}$. What frequency does the nurse observe? Assume the speed of sound is $v = 343.00\,\text{m/s}$.

Solution:

$v_\text{s} = 31.29\,\text{m/s}$
$f_\text{o} = 1.12\,\text{kHz}$

Exercise:

Problem:

The frequency of the siren of an ambulance is 900 Hz and is approaching you. You are standing on a corner and observe a frequency of 960 Hz. What is the speed of the ambulance (in mph) if the speed of sound is $v = 340.00\,\text{m/s}$?

Exercise:

Problem:

What is the minimum speed at which a source must travel toward you for you to be able to hear that its frequency is Doppler shifted? That is, what speed produces a shift of 0.300% on a day when the speed of sound is 331 m/s?

Solution:

An audible shift occurs when $\frac{f_\text{obs}}{f_\text{s}} \geq 1.003$; $f_\text{obs} = f_\text{s}\frac{v}{v - v_\text{s}} \Rightarrow \frac{f_\text{obs}}{f_\text{s}} = \frac{v}{v - v_\text{s}} \Rightarrow$
$v_\text{s} = 0.990\,\text{m/s}$

Glossary

Doppler effect
    alteration in the observed frequency of a sound due to motion of either the source or the observer

Doppler shift
    actual change in frequency due to relative motion of source and observer

Shock Waves

By the end of this section, you will be able to:

- Explain the mechanism behind sonic booms
- Describe the difference between sonic booms and shock waves
- Describe a bow wake

When discussing the Doppler effect of a moving source and a stationary observer, the only cases we considered were cases where the source was moving at speeds that were less than the speed of sound. Recall that the observed frequency for a moving source approaching a stationary observer is $f_\text{o} = f_\text{s}\left(\frac{v}{v - v_\text{s}}\right)$. As the source approaches the speed of sound, the observed frequency increases. According to the equation, if the source moves at the speed of sound, the denominator is equal to zero, implying the observed frequency is infinite. If the source moves at speeds greater than the speed of sound, the observed frequency is negative.

What could this mean? What happens when a source approaches the speed of sound? It was once argued by some scientists that such a large pressure wave would result from the constructive interference of the sound waves, that it would be impossible for a plane to exceed the speed of sound because the pressures would be great enough to destroy the airplane. But now planes routinely fly faster than the speed of sound. On July 28, 1976, Captain Eldon W. Joersz and Major George T. Morgan flew a Lockheed SR-71 Blackbird #61-7958 at 3529.60 km/h (2193.20 mi/h), which is Mach 2.85. The Mach number is the speed of the source divided by the speed of sound:

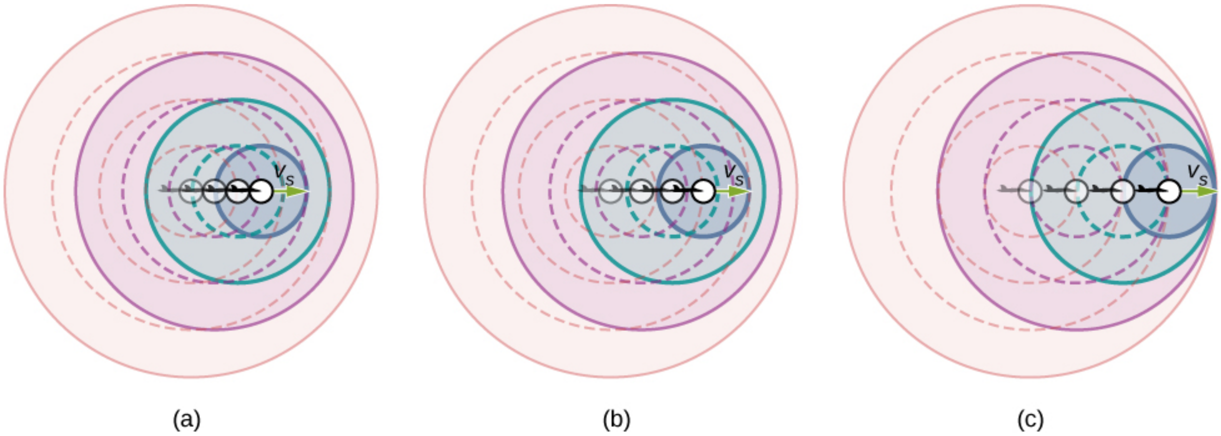
Note:
Equation:

$$M = \frac{v_\text{s}}{v}.$$

You will see that interesting phenomena occur when a source approaches and exceeds the speed of sound.

Doppler Effect and High Velocity

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer to this question applies not only to sound but to all other waves as well. Suppose a jet plane is coming nearly straight at you, emitting a sound of frequency f_s. The greater the plane's speed v_s, the greater the Doppler shift and the greater the value observed for f_o ([link]).

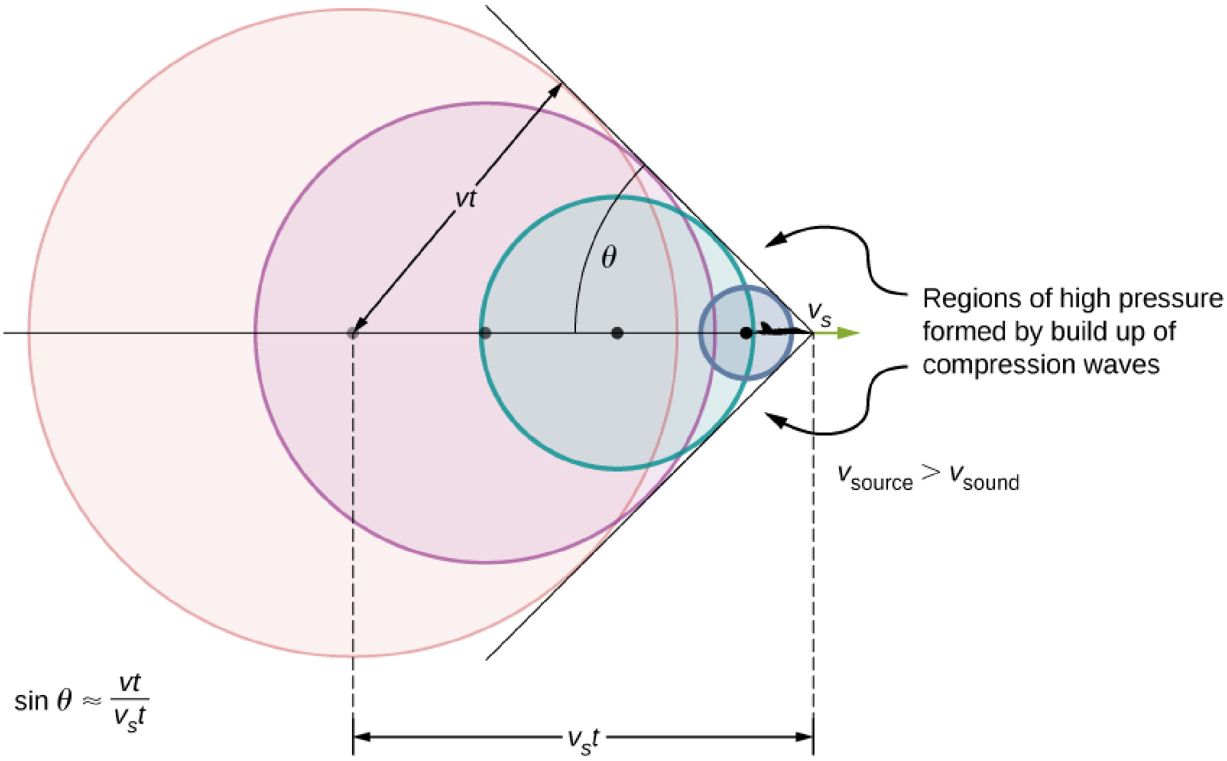


Because of the Doppler shift, as a moving source approaches a stationary observer, the observed frequency is higher than the source frequency. The faster the source is moving, the higher the observed frequency. In this figure, the source in (b) is moving faster than the source in (a). Shown are four time steps, the first three shown as dotted lines. (c) If a source moves at the speed of sound, each successive wave interfere with the previous one and the observer observes them all at the same instant.

Now, as v_s approaches the speed of sound, f_o approaches infinity, because the denominator in $f_\text{o} = f_\text{s}\left(\frac{v}{v \mp v_\text{s}}\right)$ approaches zero. At the speed of sound, this result means that in front of the source, each successive wave interferes with the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, so the frequency is infinite [part (c) of the figure].

Shock Waves and Sonic Booms

If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a shock wave is created ([link]).



vt
θ
v_s
Regions of high pressure formed by build up of compression waves
$v_\text{source} > v_\text{sound}$
$\sin\theta \approx \frac{vt}{v_s t}$
$v_s t$

Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each wave. Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

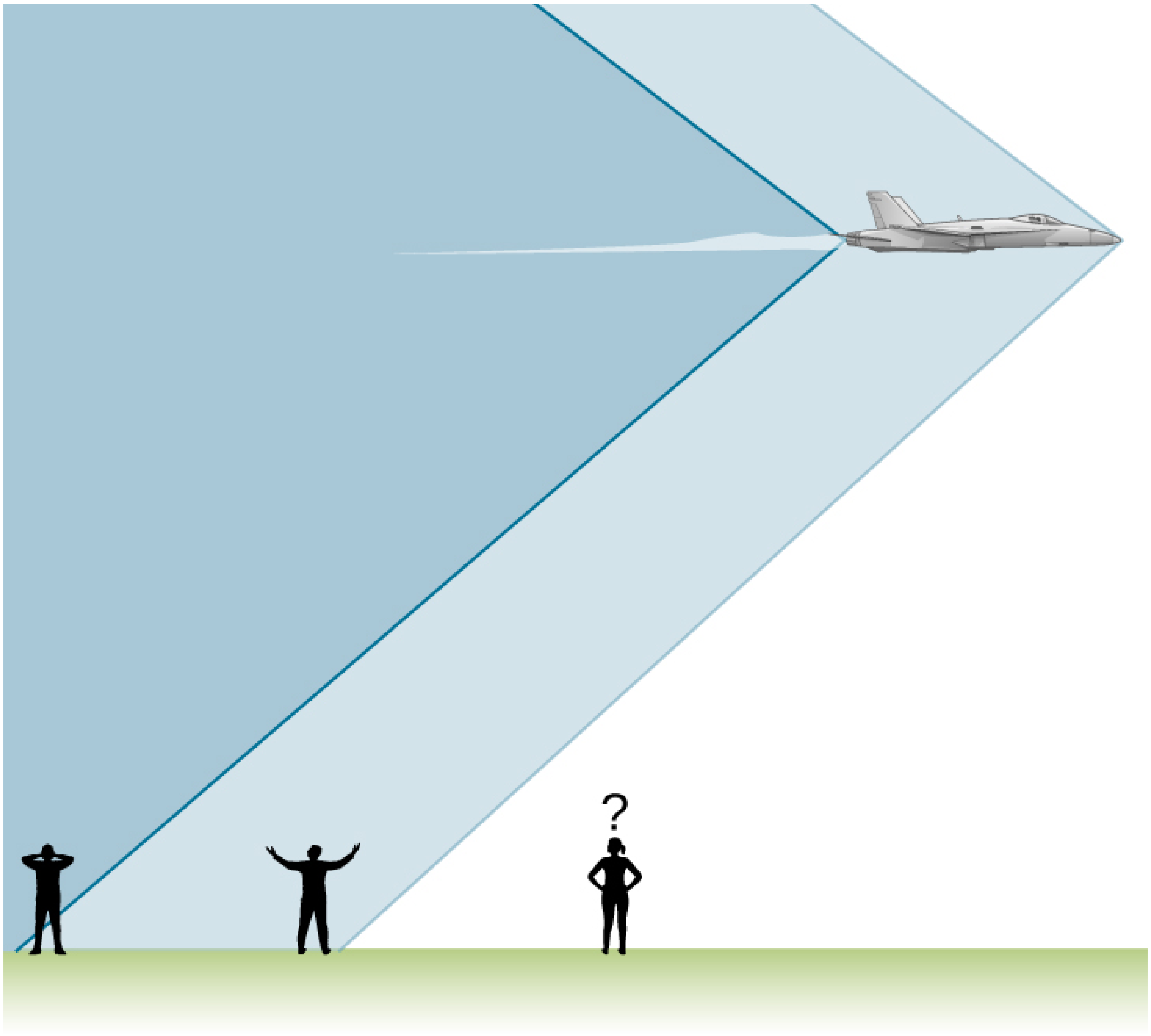
Constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **shock wave**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, so the sound intensity there is much less than on the shock wave. The angle of the shock wave can be found from the geometry. In time *t* the source has moved $v_\mathrm{s}t$ and the sound wave has moved a distance *vt* and the angle can be found using $\sin\theta = \frac{vt}{v_\mathrm{s}t} = \frac{v}{v_\mathrm{s}}$. Note that the Mach number is defined as $\frac{v_\mathrm{s}}{v}$ so the sine of the angle equals the inverse of the Mach number,

Note:
Equation:

$$\sin\theta = \frac{v}{w_\mathrm{s}} = \frac{1}{M}.$$

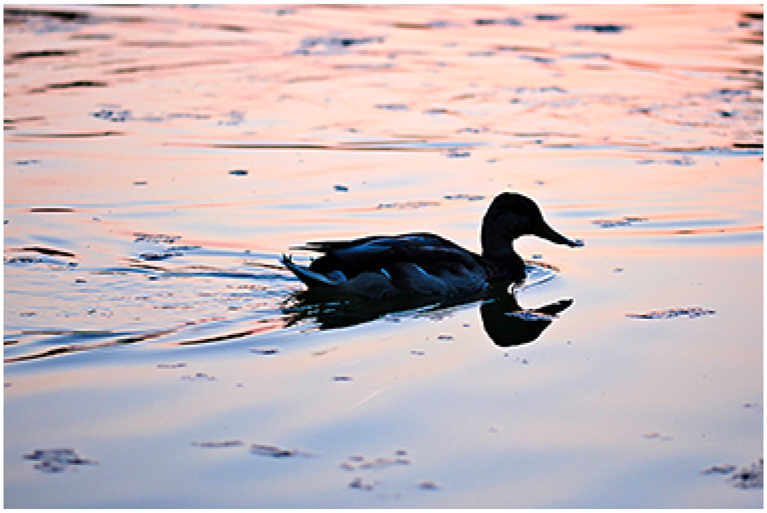
You may have heard of the common term ‘**sonic boom**.’ A common misconception is that the sonic boom occurs as the plane breaks the sound barrier; that is, accelerates to a speed higher than the speed of sound. Actually, the sonic boom occurs as the shock wave sweeps along the ground.

An aircraft creates two shock waves, one from its nose and one from its tail ([link]). During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in the figure. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas.

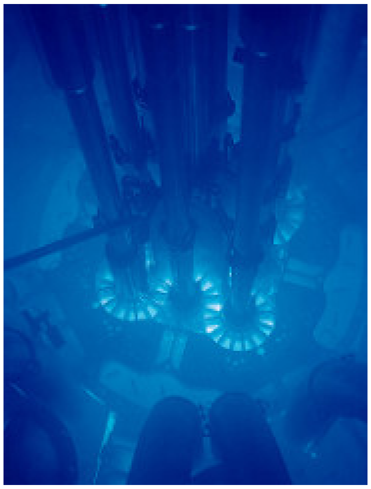


Two sonic booms experienced by observers, created by the nose and tail of an aircraft as the shock wave sweeps along the ground, are observed on the ground after the plane has passed by.

Shock waves are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in [link], is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake, trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium. (In a vacuum, the maximum speed of light is $c = 3.00 \times 10^8 \text{ m/s}$; in the medium of water, the speed of light is closer to $0.75c$.) If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in [link]. Such a bow wake is called Cerenkov radiation and is commonly observed in particle physics.



Bow wake created by a duck. Constructive interference produces the rather structured wake, whereas relatively little wave action occurs inside the wake, where interference is mostly destructive. (credit: Horia Varlan)



The blue glow in this research reactor pool is Cerenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: Idaho National Laboratory)

Summary

- The Mach number is the velocity of a source divided by the speed of sound, $M = \frac{v_s}{v}$.
- When a sound source moves faster than the speed of sound, a shock wave is produced as the sound waves interfere.

- A sonic boom is the intense sound that occurs as the shock wave moves along the ground.
- The angle the shock wave produces can be found as $\sin\theta = \frac{v}{v_s} = \frac{1}{M}$.
- A bow wake is produced when an object moves faster than the speed of a mechanical wave in the medium, such as a boat moving through the water.

Key Equations

Pressure of a sound wave	$\Delta P = \Delta P_{\text{max}} \sin\left(kx \mp \omega t + \phi\right)$
Displacement of the oscillating molecules of a sound wave	$s\left(x,t\right) = s_{\text{max}} \cos\left(kx \mp \omega t + \phi\right)$
Velocity of a wave	$v = f\lambda$
Speed of sound in a fluid	$v = \sqrt{\frac{\beta}{\rho}}$
Speed of sound in a solid	$v = \sqrt{\frac{Y}{\rho}}$
Speed of sound in an ideal gas	$v = \sqrt{\frac{\gamma RT}{M}}$
Speed of sound in air as a function of temperature	$v = 331\,\frac{\text{m}}{\text{s}}\sqrt{\frac{T_{\text{K}}}{273\,\text{K}}} = 331\,\frac{\text{m}}{\text{s}}\sqrt{1 + \frac{T_{\text{C}}}{273^\circ\text{C}}}$

Decrease in intensity as a spherical wave expands	$I_2 = I_1\left(\frac{r_1}{r_2}\right)^2$
Intensity averaged over a period	$I = \frac{\langle P \rangle}{A}$
Intensity of sound	$I = \frac{(\Delta p_{\text{max}})^2}{2\rho v}$
Sound intensity level	$\beta\,(dB) = 10\log_{10}\left(\frac{I}{I_0}\right)$
Resonant wavelengths of a tube closed at one end	$\lambda_n = \frac{4}{n}L, \;\; n = 1, 3, 5, \ldots$
Resonant frequencies of a tube closed at one end	$f_n = n\frac{v}{4L}, \;\; n = 1, 3, 5, \ldots$
Resonant wavelengths of a tube open at both ends	$\lambda_n = \frac{2}{n}L, \;\; n = 1, 2, 3, \ldots$
Resonant frequencies of a tube open at both ends	$f_n = n\frac{v}{2L}, \;\; n = 1, 2, 3, \ldots$
Beat frequency produced by two waves that differ in frequency	$f_{\text{beat}} = \lvert f_2 - f_1 \rvert$
Observed frequency for a stationary observer and a moving source	$f_{\text{o}} = f_{\text{s}}\left(\frac{v}{v \mp v_{\text{s}}}\right)$

Observed frequency for a moving observer and a stationary source	$f_{\text{o}} = f_{\text{s}} \left(\frac{v \pm v_{\text{o}}}{v} \right)$
Doppler shift for the observed frequency	$f_{\text{o}} = f_{\text{s}} \left(\frac{v \pm v_{\text{o}}}{v \mp v_{\text{s}}} \right)$
Mach number	$M = \frac{v_s}{v}$
Sine of angle formed by shock wave	$\sin\theta = \frac{v}{v_s} = \frac{1}{M}$

Conceptual Questions

Exercise:

Problem:

What is the difference between a sonic boom and a shock wave?

Exercise:

Problem:

Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

Solution:

The speed of sound decreases as the temperature decreases. The Mach number is equal to $M = \frac{v_{\text{s}}}{v}$, so the plane should slow down.

Exercise:

Problem:

When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

Problems

Exercise:

Problem:

An airplane is flying at Mach 1.50 at an altitude of 7500.00 meters, where the speed of sound is $v = 343.00\,\text{m/s}$. How far away from a stationary observer will the plane be when the observer hears the sonic boom?

Exercise:

Problem:

A jet flying at an altitude of 8.50 km has a speed of Mach 2.00, where the speed of sound is $v = 340.00\,\text{m/s}$. How long after the jet is directly overhead, will a stationary observer hear a sonic boom?

Solution:

$$\theta = 30.02^\circ$$
$$v_\text{s} = 680.00\,\text{m/s}$$
$$\tan\theta = \frac{y}{v_\text{s}t}, \quad t = 21.65\,\text{s}$$

Exercise:

Problem:

The shock wave off the front of a fighter jet has an angle of $\theta = 70.00^\circ$. The jet is flying at 1200 km/h. What is the speed of sound?

Exercise:

Problem:

A plane is flying at Mach 1.2, and an observer on the ground hears the sonic boom 15.00 seconds after the plane is directly overhead. What is the altitude of the plane? Assume the speed of sound is $v_{\text{w}} = 343.00\,\text{m/s}$.

Solution:

$\sin\theta = \frac{1}{M}, \;\; \theta = 56.47^{\circ}$
$y = 9.31\,\text{km}$

Exercise:

Problem:

A bullet is fired and moves at a speed of 1342 mph. Assume the speed of sound is $v = 340.00\,\text{m/s}$. What is the angle of the shock wave produced?

Exercise:

Problem:

A speaker is placed at the opening of a long horizontal tube. The speaker oscillates at a frequency of *f*, creating a sound wave that moves down the tube. The wave moves through the tube at a speed of $v = 340.00\,\text{m/s}$. The sound wave is modeled with the wave function $s\left(x,t\right) = s_{\text{max}}\text{cos}\left(kx - \omega t + \phi\right)$. At time $t = 0.00\,\text{s}$, an air molecule at $x = 2.3\,\text{m}$ is at the maximum displacement of 6.34 nm. At the same time, another molecule at $x = 2.7\,\text{m}$ has a displacement of 2.30 nm. What is the wave function of the sound wave, that is, find the wave number, angular frequency, and the initial phase shift?

Solution:

$$s_1 = 6.34\,\text{nm}$$
$$s_2 = 2.30\,\text{nm}$$
$$kx_1 + \phi = 0\,\text{rad}$$
$$kx_2 + \phi = 1.20\,\text{rad}$$
$$k\left(x_2 - x_1\right) = 1.20\,\text{rad}$$
$$k = 3.00\,\text{m}^{-1}$$
$$\omega = 1019.62\,\text{s}^{-1}$$
$$s_1 = s_{\text{max}}\text{cos}\left(kx_1 - \phi\right)$$
$$\phi = 5.66\,\text{rad}$$
$$s\left(x, t\right) = 6.30\,\text{nm}\text{cos}\left(3.00\,\text{m}^{-1}x - 1019.62\,\text{s}^{-1}t + 5.66\right)$$

Exercise:

Problem:

An airplane moves at Mach 1.2 and produces a shock wave. (a) What is the speed of the plane in meters per second? (b) What is the angle that the shock wave moves?

Additional Problems

Exercise:

Problem:

A 0.80-m-long tube is opened at both ends. The air temperature is 26°C. The air in the tube is oscillated using a speaker attached to a signal generator. What are the wavelengths and frequencies of first two modes of sound waves that resonate in the tube?

Solution:

$v_\text{s} = 346.40\,\text{m/s};$

$\lambda_n = \frac{2}{n}L \qquad f_n = \frac{v_\text{s}}{\lambda_n}$

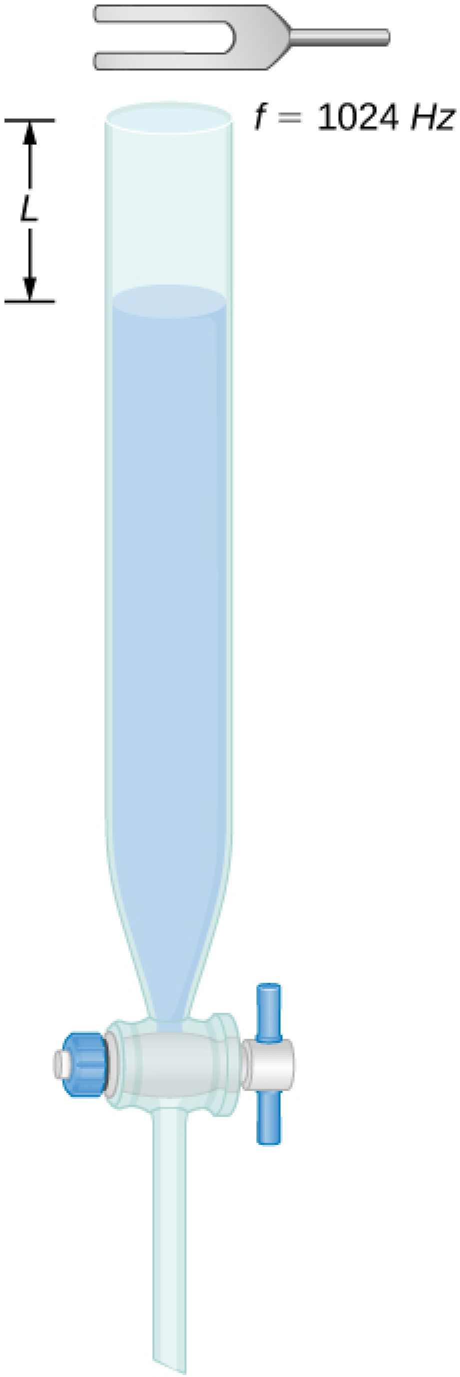
$\lambda_1 = 1.60\,\text{m} \quad f_1 = 216.50\,\text{Hz}$

$\lambda_2 = 0.80\,\text{m} \quad f_1 = 433.00\,\text{Hz}$

Exercise:

Problem:

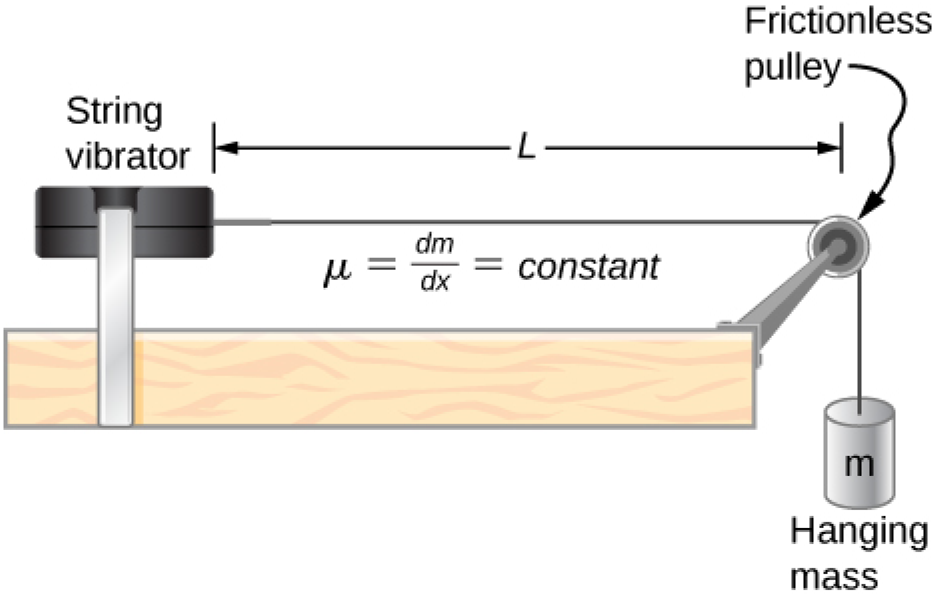
A tube filled with water has a valve at the bottom to allow the water to flow out of the tube. As the water is emptied from the tube, the length L of the air column changes. A 1024-Hz tuning fork is placed at the opening of the tube. Water is removed from the tube until the $n = 5$ mode of a sound wave resonates. What is the length of the air column if the temperature of the air in the room is 18°C?



Exercise:

Problem:

Consider the following figure. The length of the string between the string vibrator and the pulley is $L = 1.00\,\mathrm{m}$. The linear density of the string is $\mu = 0.006\,\mathrm{kg/m}$. The string vibrator can oscillate at any frequency. The hanging mass is 2.00 kg. (a)What are the wavelength and frequency of $n = 6$ mode? (b) The string oscillates the air around the string. What is the wavelength of the sound if the speed of the sound is $v_s = 343.00\,\mathrm{m/s}$?


Frictionless
pulley
String
vibrator
L
$\mu = \frac{dm}{dx} = constant$
m
Hanging
mass

Solution:

a. $$\begin{gathered}\lambda_6 = 0.40\,\mathrm{m}\\ v = 57.15\,\frac{\mathrm{m}}{\mathrm{s}}\\ f_6 = 142.89\,\mathrm{Hz}\end{gathered}$$; b. $\lambda_\mathrm{s} = 2.40\,\mathrm{m}$

Exercise:

Problem:

Early Doppler shift experiments were conducted using a band playing music on a train. A trumpet player on a moving railroad flatcar plays a 320-Hz note. The sound waves heard by a stationary observer on a train platform hears a frequency of 350 Hz. What is the flatcar's speed in mph? The temperature of the air is $T_{\text{C}} = 22^{\circ}\text{C}$.

Exercise:

Problem:

Two cars move toward one another, both sounding their horns $(f_s = 800\text{ Hz})$. Car A is moving at 65 mph and Car B is at 75 mph. What is the beat frequency heard by each driver? The air temperature is $T_C = 22.00^{\circ}\text{C}$.

Solution:

$$v = 344.08\,\frac{\text{m}}{\text{s}}$$

$$v_A = 29.05\,\frac{\text{m}}{\text{s}},\quad v_B = 33.52\text{ m/s}$$

$$f_A = 961.18\text{ Hz},$$

$$f_B = 958.89\text{ Hz}$$

$$f_{A,\text{beat}} = 161.18\text{ Hz},\quad f_{B,\text{beat}} = 158.89\text{ Hz}$$

Exercise:

Problem:

Student A runs after Student B. Student A carries a tuning fork ringing at 1024 Hz, and student B carries a tuning fork ringing at 1000 Hz. Student A is running at a speed of $v_A = 5.00\text{ m/s}$ and Student B is running at $v_B = 6.00\text{ m/s}$. What is the beat frequency heard by each student? The speed of sound is $v = 343.00\text{ m/s}$.

Exercise:

Problem:

Suppose that the sound level from a source is 75 dB and then drops to 52 dB, with a frequency of 600 Hz. Determine the (a) initial and (b) final sound intensities and the (c) initial and (d) final sound wave amplitudes. The air temperature is $T_{\text{C}} = 24.00^\circ\text{C}$ and the air density is $\rho = 1.184\,\text{kg/m}^3$.

Solution:

$v = 345.24\,\frac{\text{m}}{\text{s}}$; a. $I = 31.62\,\frac{\mu\text{W}}{\text{m}^2}$; b. $I = 0.16\,\frac{\mu\text{W}}{\text{m}^2}$; c. $s_{\text{max}} = 104.39\ \mu\text{m}$; d. $s_{\text{max}} = 7.43\ \mu\text{m}$

Exercise:

Problem:

The Doppler shift for a Doppler radar is found by $f = f_R\left(\frac{1+\frac{v}{c}}{1-\frac{v}{c}}\right)$, where f_R is the frequency of the radar, *f* is the frequency observed by the radar, *c* is the speed of light, and *v* is the speed of the target. What is the beat frequency observed at the radar, assuming the speed of the target is much slower than the speed of light?

Exercise:

Problem:

A stationary observer hears a frequency of 1000.00 Hz as a source approaches and a frequency of 850.00 Hz as a source departs. The source moves at a constant velocity of 75 mph. What is the temperature of the air?

Solution:

$$\frac{f_A}{f_D} = \frac{v+v_{\text{s}}}{v-v_{\text{s}}},\quad (v - v_{\text{s}})\frac{f_A}{f_D} = v + v_{\text{s}},\quad v = 347.39\,\frac{\text{m}}{\text{s}}$$
$$T_{\text{C}} = 27.70^\circ$$

Exercise:

Problem:

A flute plays a note with a frequency of 600 Hz. The flute can be modeled as a pipe open at both ends, where the flute player changes the length with his finger positions. What is the length of the tube if this is the fundamental frequency?

Challenge Problems

Exercise:

Problem:

Two sound speakers are separated by a distance *d*, each sounding a frequency *f*. An observer stands at one speaker and walks in a straight line a distance *x*, perpendicular to the the two speakers, until he comes to the first maximum intensity of sound. The speed of sound is *v*. How far is he from the speaker?

Solution:

$$\sqrt{x^2 + d^2} - x = \lambda, \;\; x^2 + d^2 = (\lambda + x)^2$$

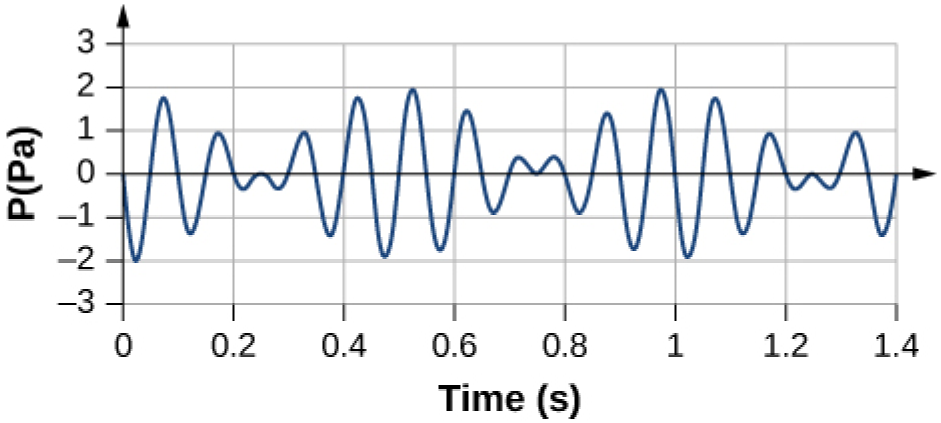
$$x^2 + d^2 = \lambda^2 + 2x\lambda + x^2, \;\; d^2 = \lambda^2 + 2x\lambda$$

$$x = \frac{d^2 - \left(\frac{v}{f}\right)^2}{2\frac{v}{f}}$$

Exercise:

Problem:

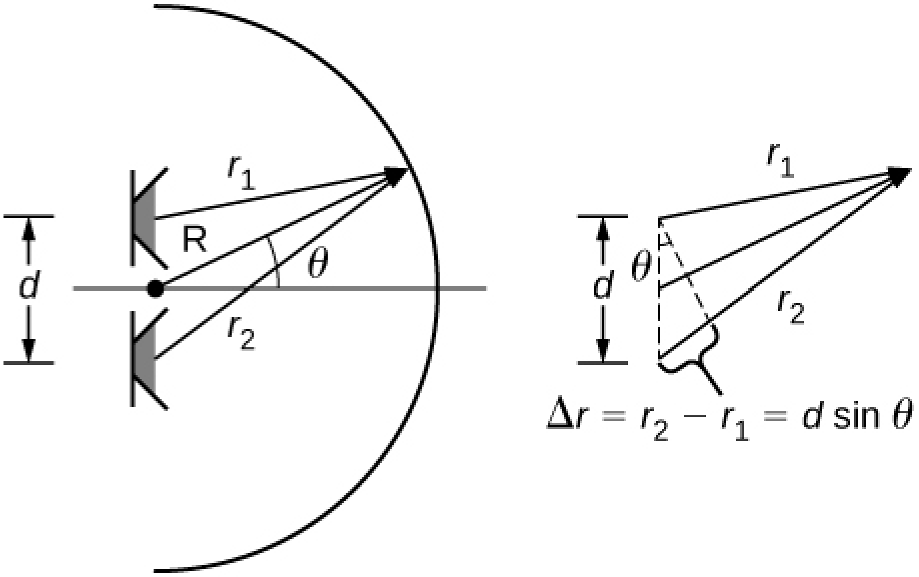
Consider the beats shown below. This is a graph of the gauge pressure versus time for the position $x = 0.00\,\text{m}$. The wave moves with a speed of $v = 343.00\,\text{m/s}$. (a) How many beats are there per second? (b) How many times does the wave oscillate per second? (c) Write a wave function for the gauge pressure as a function of time.



Exercise:

Problem:

Two speakers producing the same frequency of sound are a distance of *d* apart. Consider an arc along a circle of radius *R*, centered at the midpoint of the speakers, as shown below. (a) At what angles will there be maxima? (b) At what angle will there be minima?



Solution:

a. For maxima

$$\Delta r = d \sin \theta$$
$$d \sin \theta = n\lambda \;\; n = 0, \pm 1, \pm 2...., \;\; \theta = \sin^{-1}\left(n\frac{\lambda}{d}\right) \;\; n = 0, \pm 1, \pm 2....$$

b. For minima,
$$\begin{aligned} &\Delta r = d \sin \theta \\ &d \sin \theta = \left(n + \tfrac{1}{2}\right)\lambda \;\; n = 0, \pm 1, \pm 2.... \\ &\theta = \sin^{-1}\left(\left(n + \tfrac{1}{2}\right)\frac{\lambda}{d}\right) \;\; n = 0, \pm 1, \pm 2.... \end{aligned}$$

Exercise:

Problem:

A string has a length of 1.5 m, a linear mass density $\mu = 0.008 \, \text{kg/m}$, , and a tension of 120 N. If the air temperature is $T = 22^\circ\text{C}$, what should the length of a pipe open at both ends for it to have the same frequency for the $n = 3$ mode?

Exercise:

Problem:

A string $\left(\mu = 0.006\,\frac{\text{kg}}{\text{m}}, L = 1.50\text{ m}\right)$ is fixed at both ends and is under a tension of 155 N. It oscillates in the $n = 10$ mode and produces sound. A tuning fork is ringing nearby, producing a beat frequency of 23.76 Hz. (a) What is the frequency of the sound from the string? (b) What is the frequency of the tuning fork if the tuning fork frequency is lower? (c) What should be the tension of the string for the beat frequency to be zero?

Solution:

a. $v_{\text{string}} = 160.73\,\frac{\text{m}}{\text{s}}, \quad f_{\text{string}} = 535.77\text{ Hz}$; b. $f_{\text{fork}} = 512\text{ Hz}$; c. $f_{\text{fork}} = \frac{n\sqrt{\frac{F_T}{\mu}}}{2L}, \quad F_T = 141.56\text{ N}$

Exercise:

Problem:

A string has a linear mass density μ, a length L, and a tension of F_T, and oscillates in a mode *n* at a frequency *f*. Find the ratio of $\frac{\Delta f}{f}$ for a small change in tension.

Exercise:

Problem:

A string has a linear mass density $\mu = 0.007\text{ kg/m}$, a length $L = 0.70\text{ m}$, a tension of $F_T = 110\text{ N}$, and oscillates in a mode $n = 3$. (a) What is the frequency of the oscillations? (b) Use the result in the preceding problem to find the change in the frequency when the tension is increased by 1.00%.

Solution:

a. $f = 268.62\text{ Hz}$; b. $\Delta f \approx \frac{1}{2}\,\frac{\Delta F_T}{F_T}\,f = 1.34\text{ Hz}$

Exercise:

Problem:

A speaker powered by a signal generator is used to study resonance in a tube. The signal generator can be adjusted from a frequency of 1000 Hz to 1800 Hz. First, a 0.75-m-long tube, open at both ends, is studied. The temperature in the room is $T_{\text{F}} = 85.00^{\circ}\text{F}$. (a) Which normal modes of the pipe can be studied? What are the frequencies and wavelengths? Next a cap is place on one end of the 0.75-meter-long pipe. (b) Which normal modes of the pipe can be studied? What are the frequencies and wavelengths?

Exercise:

Problem:

A string on the violin has a length of 23.00 cm and a mass of 0.900 grams. The tension in the string 850.00 N. The temperature in the room is $T_C = 24.00^{\circ}\text{C}$. The string is plucked and oscillates in the $n = 9$ mode. (a) What is the speed of the wave on the string? (b) What is the wavelength of the sounding wave produced? (c) What is the frequency of the oscillating string? (d) What is the frequency of the sound produced? (e) What is the wavelength of the sound produced?

Solution:

a. $v = 466.07\,\frac{\text{m}}{\text{s}}$; b. $\lambda_9 = 51.11\,\text{mm}$; c. $f_9 = 9.12\,\text{kHz}$; d. $f_{\text{sound}} = 9.12\,\text{kHz}$; e. $\lambda_{\text{air}} = 37.86\,\text{mm}$

Glossary

bow wake	v-shaped disturbance created when the wave source moves faster than the wave propagation speed
shock wave	

wave front that is produced when a sound source moves faster than the speed of sound

sonic boom

loud noise that occurs as a shock wave as it sweeps along the ground

Introduction class="introduction"

Electric charges exist all around us. They can cause objects to be repelled from each other or to be attracted to each other. (credit: modification of work by Sean McGrath)

Back when we were studying Newton’s laws, we identified several physical phenomena as forces. We did so based on the effect they had on a physical object: Specifically, they caused the object to accelerate. Later, when we studied impulse and momentum, we expanded this idea to identify a force as any physical phenomenon that changed the momentum of an object. In either case, the result is the same: We recognize a force by the effect that it has on an object.

In Gravitation, we examined the force of gravity, which acts on all objects with mass. In this chapter, we begin the study of the electric force, which acts on all objects with a property called charge. The electric force is much stronger than gravity (in most systems where both appear), but it can be a force of attraction or a force of repulsion, which leads to very different effects on objects. The electric force helps keep atoms together, so it is of fundamental importance in matter. But it also governs most everyday interactions we deal with, from chemical interactions to biological processes.

Electric Charge

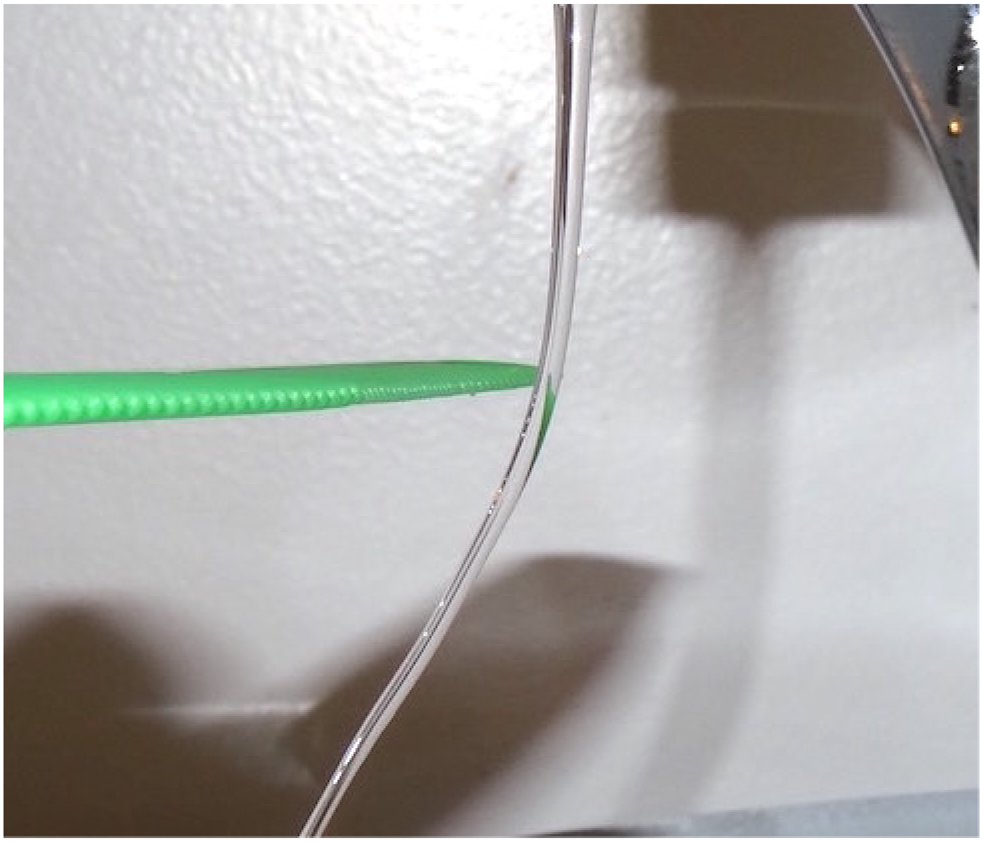
By the end of this section, you will be able to:

- Describe the concept of electric charge
- Explain qualitatively the force electric charge creates

You are certainly familiar with electronic devices that you activate with the click of a switch, from computers to cell phones to television. And you have certainly seen electricity in a flash of lightning during a heavy thunderstorm. But you have also most likely experienced electrical effects in other ways, maybe without realizing that an electric force was involved. Let’s take a look at some of these activities and see what we can learn from them about electric charges and forces.

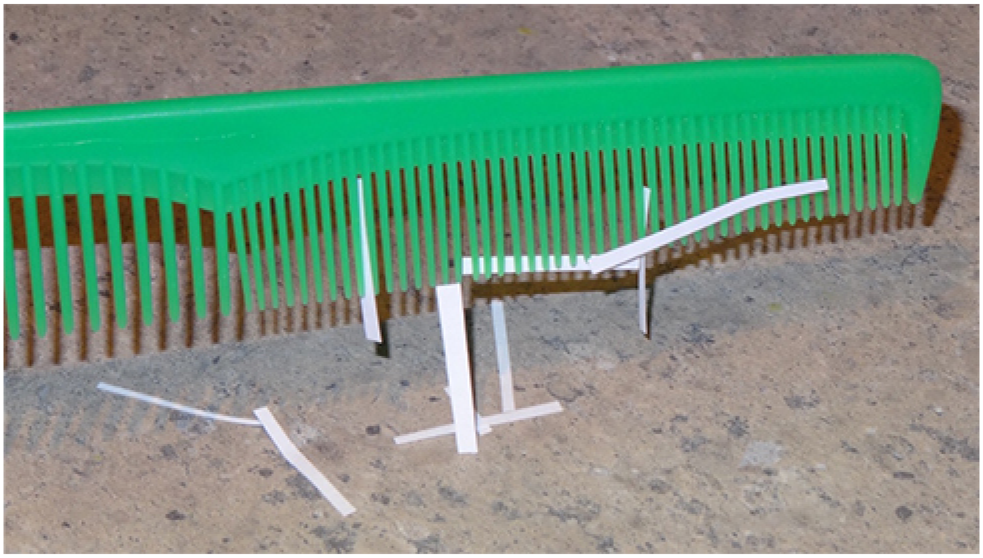
Discoveries

You have probably experienced the phenomenon of **static electricity**: When you first take clothes out of a dryer, many (not all) of them tend to stick together; for some fabrics, they can be very difficult to separate. Another example occurs if you take a woolen sweater off quickly—you can feel (and hear) the static electricity pulling on your clothes, and perhaps even your hair. If you comb your hair on a dry day and then put the comb close to a thin stream of water coming out of a faucet, you will find that the water stream bends toward (is attracted to) the comb ([link]).



An electrically charged comb attracts a stream of water from a distance. Note that the water is not touching the comb. (credit: Jane Whitney)

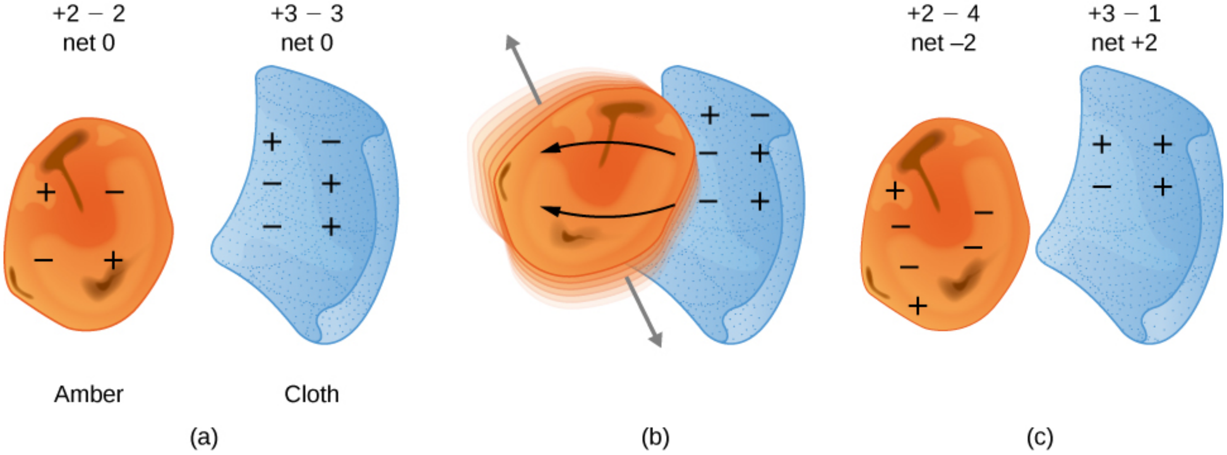
Suppose you bring the comb close to some small strips of paper; the strips of paper are attracted to the comb and even cling to it ([link]). In the kitchen, quickly pull a length of plastic cling wrap off the roll; it will tend to cling to most any nonmetallic material (such as plastic, glass, or food). If you rub a balloon on a wall for a few seconds, it will stick to the wall. Probably the most annoying effect of static electricity is getting shocked by a doorknob (or a friend) after shuffling your feet on some types of carpeting.



After being used to comb hair, this comb attracts small strips of paper from a distance, without physical contact. Investigation of this behavior helped lead to the concept of the electric force. (credit: Jane Whitney)

Many of these phenomena have been known for centuries. The ancient Greek philosopher Thales of Miletus (624–546 BCE) recorded that when amber (a hard, translucent, fossilized resin from extinct trees) was vigorously rubbed with a piece of fur, a force was created that caused the fur and the amber to be attracted to each other ([link]). Additionally, he found that the rubbed amber would not only attract the fur, and the fur attract the amber, but they both could affect other (nonmetallic) objects, even if not in contact with those objects ([link]).

Borneo amber is mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of fur, the amber gains more electrons, giving it a net negative charge. At the same time, the fur, having lost electrons, becomes positively charged. (credit: “Sebakoamber”/Wikimedia Commons)


+2 − 2
net 0
+3 − 3
net 0
Amber
Cloth
(a)
(b)
+2 − 4
net –2
+3 − 1
net +2
(c)

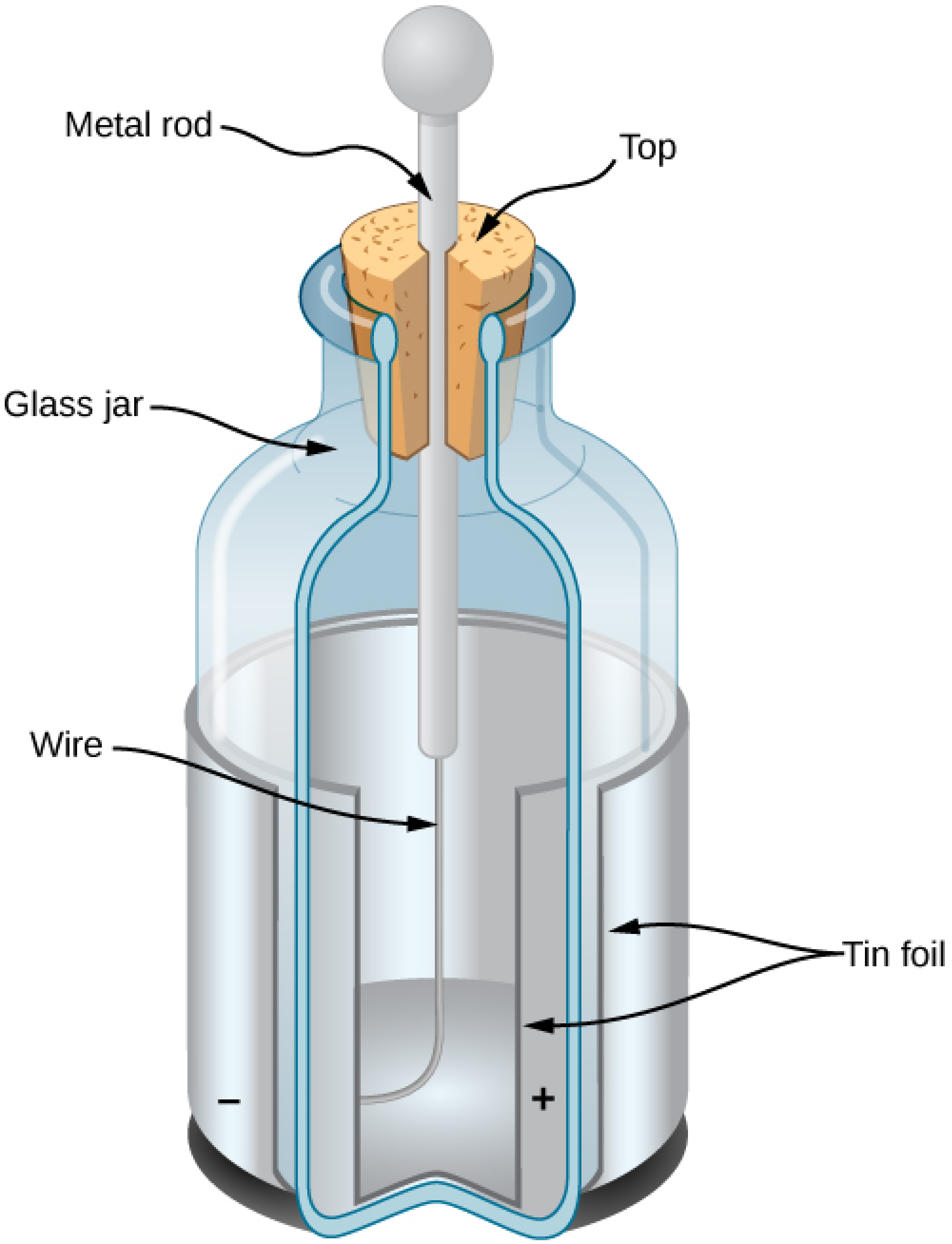
When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

The English physicist William Gilbert (1544–1603) also studied this attractive force, using various substances. He worked with amber, and, in addition, he experimented with rock crystal and various precious and semi-precious gemstones. He also experimented with several metals. He found that the metals never exhibited this force, whereas the minerals did. Moreover, although an electrified amber rod would attract a piece of fur, it would repel another electrified amber rod; similarly, two electrified pieces of fur would repel each other.

This suggested there were two types of an electric property; this property eventually came to be called **electric charge**. The difference between the two types of electric charge is in the directions of the electric forces that each type of charge causes: These forces are repulsive when the same type of charge exists on two interacting objects and attractive when the charges are of opposite types. The SI unit of electric charge is the **coulomb** (C), after the French physicist Charles-Augustin de Coulomb (1736–1806).

The most peculiar aspect of this new force is that it does not require physical contact between the two objects in order to cause an acceleration. This is an example of a so-called “long-range” force. (Or, as James Clerk Maxwell later phrased it, “action at a distance.”) With the exception of gravity, all other forces we have discussed so far act only when the two interacting objects actually touch.

The American physicist and statesman Benjamin Franklin found that he could concentrate charge in a “Leyden jar,” which was essentially a glass jar with two sheets of metal foil, one inside and one outside, with the glass between them ([link]). This created a large electric force between the two foil sheets.


Metal rod
Top
Glass jar
Wire
Tin foil
–
+

A Leyden jar (an early version of what is now called a capacitor) allowed experimenters to store large amounts of electric charge. Benjamin Franklin used such a jar to demonstrate that lightning behaved exactly like the electricity he got from the equipment in his laboratory.

Franklin pointed out that the observed behavior could be explained by supposing that one of the two types of charge remained motionless, while the other type of charge flowed from one piece of foil to the other. He further suggested that an excess of what he called this “electrical fluid” be called “positive electricity” and the deficiency of it be called “negative electricity.” His suggestion, with some minor modifications, is the model we use today. (With the experiments that he was able to do, this was a pure guess; he had no way of actually determining the sign of the moving charge. Unfortunately, he guessed wrong; we now know that the charges that flow are the ones Franklin labeled negative, and the positive charges remain largely motionless. Fortunately, as we’ll see, it makes no practical or theoretical difference which choice we make, as long as we stay consistent with our choice.)

Let’s list the specific observations that we have of this **electric force**:

- The force acts without physical contact between the two objects.
- The force can be either attractive or repulsive: If two interacting objects carry the same sign of charge, the force is repulsive; if the charges are of opposite sign, the force is attractive. These interactions are referred to as **electrostatic repulsion** and **electrostatic attraction**, respectively.
- Not all objects are affected by this force.
- The magnitude of the force decreases (rapidly) with increasing separation distance between the objects.

To be more precise, we find experimentally that the magnitude of the force decreases as the square of the distance between the two interacting objects increases. Thus, for example, when the distance between two interacting objects is doubled, the force between them decreases to one fourth what it was in the original system. We can also observe that the surroundings of the charged objects affect the magnitude of the force. However, we will explore this issue in a later chapter.

Properties of Electric Charge

In addition to the existence of two types of charge, several other properties of charge have been discovered.

- **Charge is quantized.** This means that electric charge comes in discrete amounts, and there is a smallest possible amount of charge that an object can have. In the SI system, this smallest amount is $e \equiv 1.602 \times 10^{-19}$ C. No free particle can have less charge than this, and, therefore, the charge on any object—the charge on all objects—must be an integer multiple of this amount. All macroscopic, charged objects have charge because electrons have either been added or taken away from them, resulting in a net charge.
- **The magnitude of the charge is independent of the type.** Phrased another way, the smallest possible positive charge (to four significant figures) is $+1.602 \times 10^{-19}$ C, and the smallest possible negative charge is -1.602×10^{-19} C; these values are exactly equal. This is simply how the laws of physics in our universe turned out.
- **Charge is conserved.** Charge can neither be created nor destroyed; it can only be transferred from place to place, from one object to another. Frequently, we speak of two charges “canceling”; this is verbal shorthand. It means that if two objects that have equal and opposite charges are physically close to each other, then the (oppositely directed) forces they apply on some other charged object cancel, for a net force of zero. It is important that you understand that the charges on the objects by no means disappear, however. The net charge of the universe is constant.
- **Charge is conserved in closed systems.** In principle, if a negative charge disappeared from your lab bench and reappeared on the Moon, conservation of charge would still hold. However, this never happens. If the total charge you have in your local system on your lab bench is changing, there will be a measurable flow of charge into or out of the

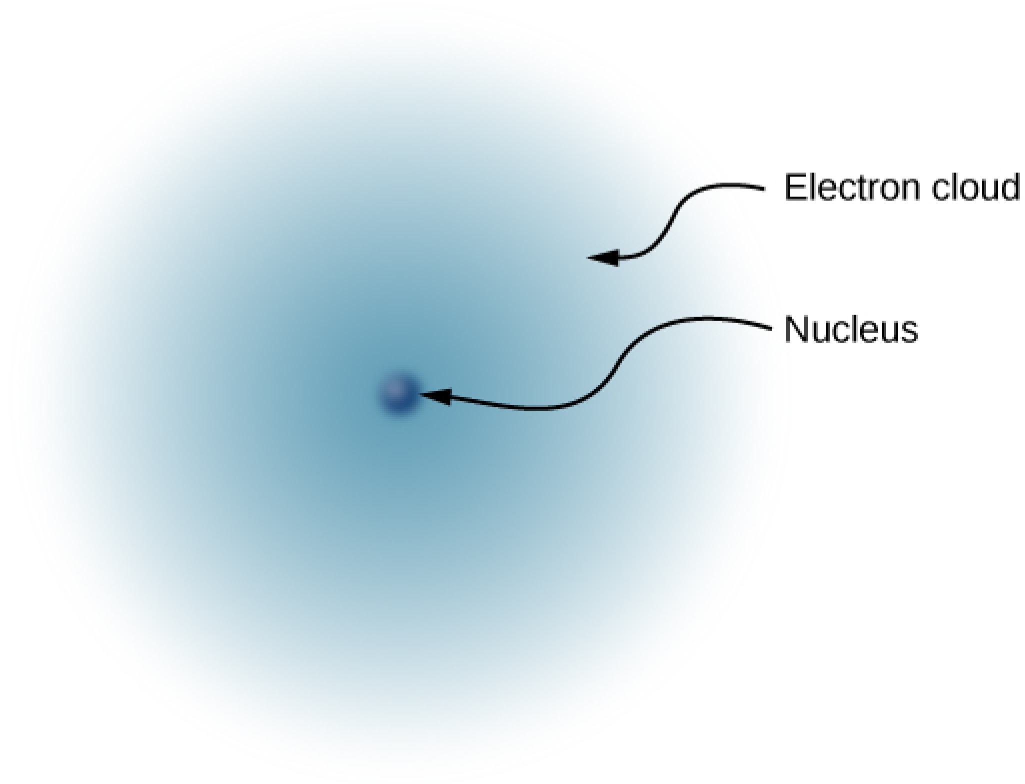
system. Again, charges can and do move around, and their effects can and do cancel, but the net charge in your local environment (if closed) is conserved. The last two items are both referred to as the **law of conservation of charge**.

The Source of Charges: The Structure of the Atom

Once it became clear that all matter was composed of particles that came to be called atoms, it also quickly became clear that the constituents of the atom included both positively charged particles and negatively charged particles. The next question was, what are the physical properties of those electrically charged particles?

The negatively charged particle was the first one to be discovered. In 1897, the English physicist J. J. Thomson was studying what was then known as *cathode rays*. Some years before, the English physicist William Crookes had shown that these "rays" were negatively charged, but his experiments were unable to tell any more than that. (The fact that they carried a negative electric charge was strong evidence that these were not rays at all, but particles.) Thomson prepared a pure beam of these particles and sent them through crossed electric and magnetic fields, and adjusted the various field strengths until the net deflection of the beam was zero. With this experiment, he was able to determine the charge-to-mass ratio of the particle. This ratio showed that the mass of the particle was much smaller than that of any other previously known particle—1837 times smaller, in fact. Eventually, this particle came to be called the **electron**.

Since the atom as a whole is electrically neutral, the next question was to determine how the positive and negative charges are distributed within the atom. Thomson himself imagined that his electrons were embedded within a sort of positively charged paste, smeared out throughout the volume of the atom. However, in 1908, the New Zealand physicist Ernest Rutherford showed that the positive charges of the atom existed within a tiny core—called a nucleus—that took up only a very tiny fraction of the overall volume of the atom, but held over 99% of the mass. (See Linear Momentum and Collisions.) In addition, he showed that the negatively charged electrons perpetually orbited about this nucleus, forming a sort of electrically charged cloud that surrounds the nucleus ([link]). Rutherford concluded that the nucleus was constructed of small, massive particles that he named **protons**.

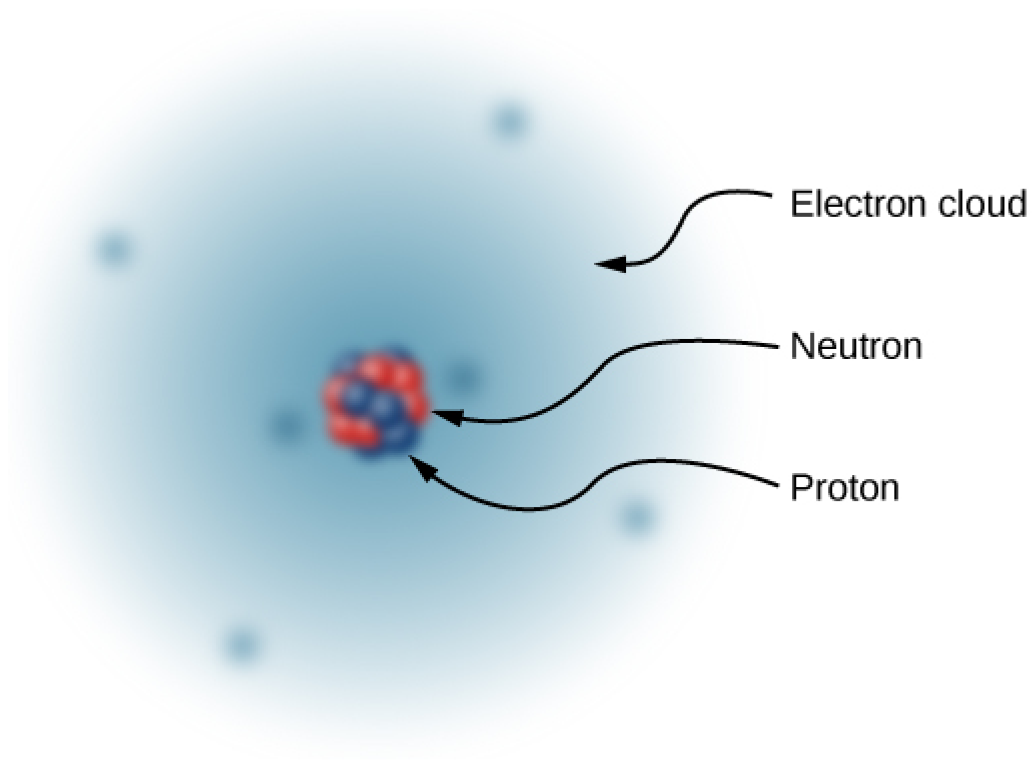


This simplified model of a hydrogen atom shows a positively charged nucleus (consisting, in the case of hydrogen, of a single proton), surrounded by an electron “cloud.” The charge of the electron cloud is equal (and opposite in sign) to the charge of the nucleus, but the electron does not have a definite location in space; hence, its representation here is as a cloud. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules, and, hence, even greater numbers of individual negative and positive charges.

Since it was known that different atoms have different masses, and that ordinarily atoms are electrically neutral, it was natural to suppose that different atoms have different numbers of protons in their nucleus, with an equal number of negatively charged electrons orbiting about the positively charged nucleus, thus making the atoms overall electrically neutral. However, it was soon discovered that although the lightest atom, hydrogen, did indeed have a single proton as its nucleus, the next heaviest atom—helium—has twice the number of protons (two), but *four* times the mass of hydrogen.

This mystery was resolved in 1932 by the English physicist James Chadwick, with the discovery of the **neutron**. The neutron is, essentially, an electrically neutral twin of the proton, with no electric charge, but (nearly) identical mass to the proton. The helium nucleus therefore has two neutrons along with its two protons. (Later experiments were to show that although the neutron is electrically neutral overall, it does have an internal charge *structure*. Furthermore, although the masses of the neutron and the proton are *nearly* equal, they aren't exactly equal: The neutron's mass is very slightly larger than the mass of the proton. That slight mass excess turned out to be of great importance. That, however, is a story that will have to wait until our study of modern physics in Nuclear Physics.)

Thus, in 1932, the picture of the atom was of a small, massive nucleus constructed of a combination of protons and neutrons, surrounded by a collection of electrons whose combined motion formed a sort of negatively charged "cloud" around the nucleus ([link]). In an electrically neutral atom, the total negative charge of the collection of electrons is equal to the total positive charge in the nucleus. The very low-mass electrons can be more or less easily removed or added to an atom, changing the net charge on the atom (though without changing its type). An atom that has had the charge altered in this way is called an **ion**. Positive ions have had electrons removed, whereas negative ions have had excess electrons added. We also use this term to describe molecules that are not electrically neutral.



The nucleus of a carbon atom is composed of six protons and six neutrons. As in hydrogen, the surrounding six electrons do not have definite locations and so can be considered to be a sort of cloud surrounding the nucleus.

The story of the atom does not stop there, however. In the latter part of the twentieth century, many more subatomic particles were discovered in the nucleus of the atom: pions, neutrinos, and quarks, among others. With the exception of the photon, none of these particles are directly relevant to the study of electromagnetism, so we defer further discussion of them until the chapter on particle physics (Particle Physics and Cosmology).

A Note on Terminology

As noted previously, electric charge is a property that an object can have. This is similar to how an object can have a property that we call mass, a property that we call density, a property that we call temperature, and so on.

Technically, we should always say something like, “Suppose we have a particle that carries a charge of $3\ \mu\text{C}$.” However, it is very common to say instead, “Suppose we have a $3\text{-}\mu\text{C}$ charge.” Similarly, we often say something like, “Six charges are located at the vertices of a regular hexagon.” A charge is not a particle; rather, it is a *property* of a particle. Nevertheless, this terminology is extremely common (and is frequently used in this book, as it is everywhere else). So, keep in the back of your mind what we really mean when we refer to a “charge.”

Summary

- There are only two types of charge, which we call positive and negative. Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- The vast majority of positive charge in nature is carried by protons, whereas the vast majority of negative charge is carried by electrons. The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge is $e \equiv 1.602\ \times\ 10^{-19}\,\text{C}$
- Both positive and negative charges exist in neutral objects and can be separated by bringing the two objects into physical contact; rubbing the objects together can remove electrons from the bonds in one object and place them on the other object, increasing the charge separation.
- For macroscopic objects, negatively charged means an excess of electrons and positively charged means a depletion of electrons.
- The law of conservation of charge states that the net charge of a closed system is constant.

Conceptual Questions

Exercise:

Problem:

There are very large numbers of charged particles in most objects. Why, then, don’t most objects exhibit static electricity?

Solution:

There are mostly equal numbers of positive and negative charges present, making the object electrically neutral.

Exercise:

Problem:

Why do most objects tend to contain nearly equal numbers of positive and negative charges?

Exercise:

Problem:

A positively charged rod attracts a small piece of cork. (a) Can we conclude that the cork is negatively charged? (b) The rod repels another small piece of cork. Can we conclude that this piece is positively charged?

Solution:

a. yes; b. yes

Exercise:

Problem:

Two bodies attract each other electrically. Do they both have to be charged? Answer the same question if the bodies repel one another.

Exercise:

Problem:

How would you determine whether the charge on a particular rod is positive or negative?

Solution:

Take an object with a known charge, either positive or negative, and bring it close to the rod. If the known charged object is positive and it is repelled from the rod, the rod is charged positive. If the positively charged object is attracted to the rod, the rod is negatively charged.

Problems

Exercise:

Problem:

Common static electricity involves charges ranging from nanocoulombs to microcoulombs. (a) How many electrons are needed to form a charge of −2.00 nC? (b) How many electrons must be removed from a neutral object to leave a net charge of $0.500\ \mu\text{C}$?

Solution:

a. $2.00\ \times\ 10^{-9}\,\text{C}\left(\frac{1}{1.602\times10^{-19}}\ \text{e/C}\right)=1.248\ \times\ 10^{10}\,\text{electrons;}$

b. $0.500\ \times\ 10^{-6}\,\text{C}\left(\frac{1}{1.602\times10^{-19}}\ \text{e/C}\right)=3.121\ \times\ 10^{12}\,\text{electrons}$

Exercise:

Problem:

If $1.80\ \times\ 10^{20}$ electrons move through a pocket calculator during a full day’s operation, how many coulombs of charge moved through it?

Exercise:

Problem:

To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?

Solution:

$\frac{3.750 \times 10^{21}\ \text{e}}{6.242 \times 10^{18}\ \text{e/C}} = –600.8\ \text{C}$

Exercise:

Problem:

A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge is this?

Exercise:

Problem:

A 2.5-g copper penny is given a charge of $-2.0 \times 10^{-9}\ \text{C}$. (a) How many excess electrons are on the penny? (b) By what percent do the excess electrons change the mass of the penny?

Solution:

a. $2.0 \times 10^{-9}\ \text{C}\ (6.242 \times 10^{18}\ \text{e/C}) = 1.248 \times 10^{10}\ \text{e}$;

b. $9.109 \times 10^{-31}\ \text{kg}\ (1.248 \times 10^{10}\ \text{e}) = 1.137 \times 10^{-20}\ \text{kg}$, $\frac{1.137 \times 10^{-20}\ \text{kg}}{2.5 \times 10^{-3}\ \text{kg}} = 4.548 \times 10^{-18}$ or $4.545 \times 10^{-16}\%$

Exercise:

Problem:

A 2.5-g copper penny is given a charge of 4.0×10^{-9} C. (a) How many electrons are removed from the penny? (b) If no more than one electron is removed from an atom, what percent of the atoms are ionized by this charging process?

Glossary

coulomb
SI unit of electric charge

electric charge
physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electric force

electric force
noncontact force observed between electrically charged objects

electron
particle surrounding the nucleus of an atom and carrying the smallest unit of negative charge

electrostatic attraction
phenomenon of two objects with opposite charges attracting each other

electrostatic repulsion
phenomenon of two objects with like charges repelling each other

ion
atom or molecule with more or fewer electrons than protons

law of conservation of charge
net electric charge of a closed system is constant

neutron

neutral particle in the nucleus of an atom, with (nearly) the same mass as a proton

proton

particle in the nucleus of an atom and carrying a positive charge equal in magnitude to the amount of negative charge carried by an electron

static electricity

buildup of electric charge on the surface of an object; the arrangement of the charge remains constant (“static”)

Conductors, Insulators, and Charging by Induction

By the end of this section, you will be able to:

- Explain what a conductor is
- Explain what an insulator is
- List the differences and similarities between conductors and insulators
- Describe the process of charging by induction

In the preceding section, we said that scientists were able to create electric charge only on nonmetallic materials and never on metals. To understand why this is the case, you have to understand more about the nature and structure of atoms. In this section, we discuss how and why electric charges do—or do not—move through materials ([link]). A more complete description is given in a later chapter.

Computer jack
Plug
Power cord

This power adapter uses metal wires and connectors to conduct electricity from the wall socket to a laptop computer. The conducting wires allow electrons to move freely through the cables, which are shielded by rubber and plastic. These materials act as insulators that don't allow electric charge to escape outward. (credit: modification of work by "Evan-Amos"/Wikimedia Commons)

Conductors and Insulators

As discussed in the previous section, electrons surround the tiny nucleus in the form of a (comparatively) vast cloud of negative charge. However, this cloud does have a definite structure to it. Let's consider an atom of the most commonly used conductor, copper.

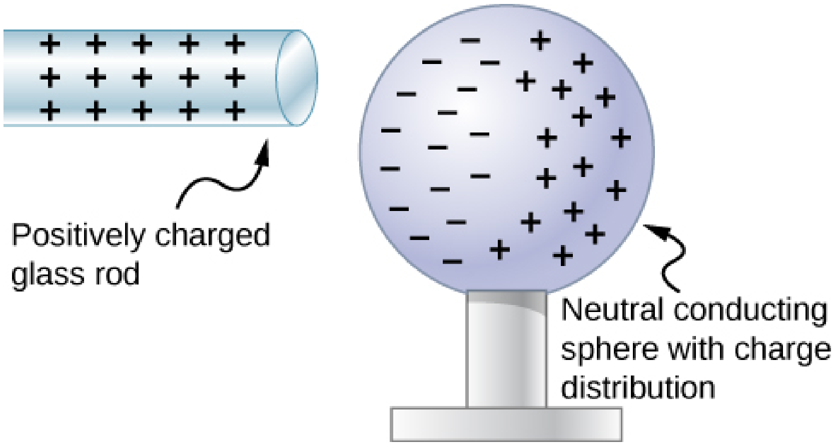
For reasons that will become clear in Atomic Structure, there is an outermost electron that is only loosely bound to the atom’s nucleus. It can be easily dislodged; it then moves to a neighboring atom. In a large mass of copper atoms (such as a copper wire or a sheet of copper), these vast numbers of outermost electrons (one per atom) wander from atom to atom, and are the electrons that do the moving when electricity flows. These wandering, or “free,” electrons are called **conduction electrons**, and copper is therefore an excellent **conductor** (of electric charge). All conducting elements have a similar arrangement of their electrons, with one or two conduction electrons. This includes most metals.

Insulators, in contrast, are made from materials that lack conduction electrons; charge flows only with great difficulty, if at all. Even if excess charge is added to an insulating material, it cannot move, remaining indefinitely in place. This is why insulating materials exhibit the electrical attraction and repulsion forces described earlier, whereas conductors do not; any excess charge placed on a conductor would instantly flow away (due to mutual repulsion from existing charges), leaving no excess charge around to create forces. Charge cannot flow along or through an **insulator**, so its electric forces remain for long periods of time. (Charge will dissipate from an insulator, given enough time.) As it happens, amber, fur, and most semi-precious gems are insulators, as are materials like wood, glass, and plastic.

Charging by Induction

Let’s examine in more detail what happens in a conductor when an electrically charged object is brought close to it. As mentioned, the conduction electrons in the conductor are able to move with nearly complete freedom. As a result, when a charged insulator (such as a positively charged glass rod) is brought close to the conductor, the (total) charge on the insulator exerts an electric force on the conduction electrons. Since the rod is positively charged, the conduction electrons (which themselves are negatively charged) are attracted, flowing toward the insulator to the near side of the conductor ([link]).

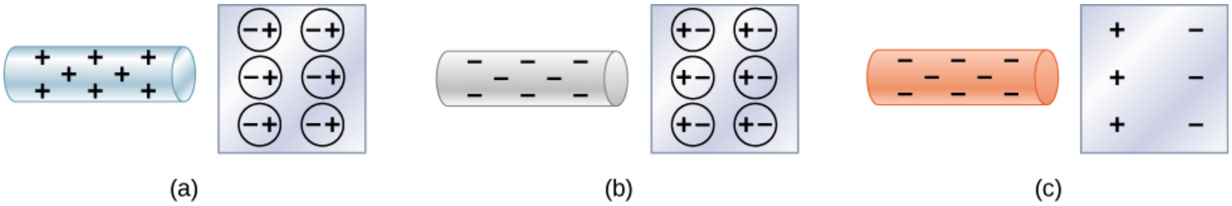
Now, the conductor is still overall electrically neutral; the conduction electrons have changed position, but they are still in the conducting material. However, the conductor now has a charge *distribution*; the near end (the portion of the conductor closest to the insulator) now has more negative charge than positive charge, and the reverse is true of the end farthest from the insulator. The relocation of negative charges to the near side of the conductor results in an overall positive charge in the part of the conductor farthest from the insulator. We have thus created an electric charge distribution where one did not exist before. This process is referred to as *inducing polarization*—in this case, polarizing the conductor. The resulting separation of positive and negative charge is called **polarization**, and a material, or even a molecule, that exhibits polarization is said to be polarized. A similar situation occurs with a negatively charged insulator, but the resulting polarization is in the opposite direction.


Positively charged glass rod
Neutral conducting sphere with charge distribution

Induced polarization. A positively charged glass rod is brought near the left side of the conducting sphere, attracting negative charge and leaving the other side of the sphere positively charged. Although the sphere is overall still electrically neutral, it now has a charge distribution, so it can exert an electric force on other nearby charges. Furthermore, the distribution is such that it will be attracted to the glass rod.

The result is the formation of what is called an electric **dipole**, from a Latin phrase meaning “two ends.” The presence of electric charges on the insulator—and the electric forces they apply to the conduction electrons—creates, or “induces,” the dipole in the conductor.

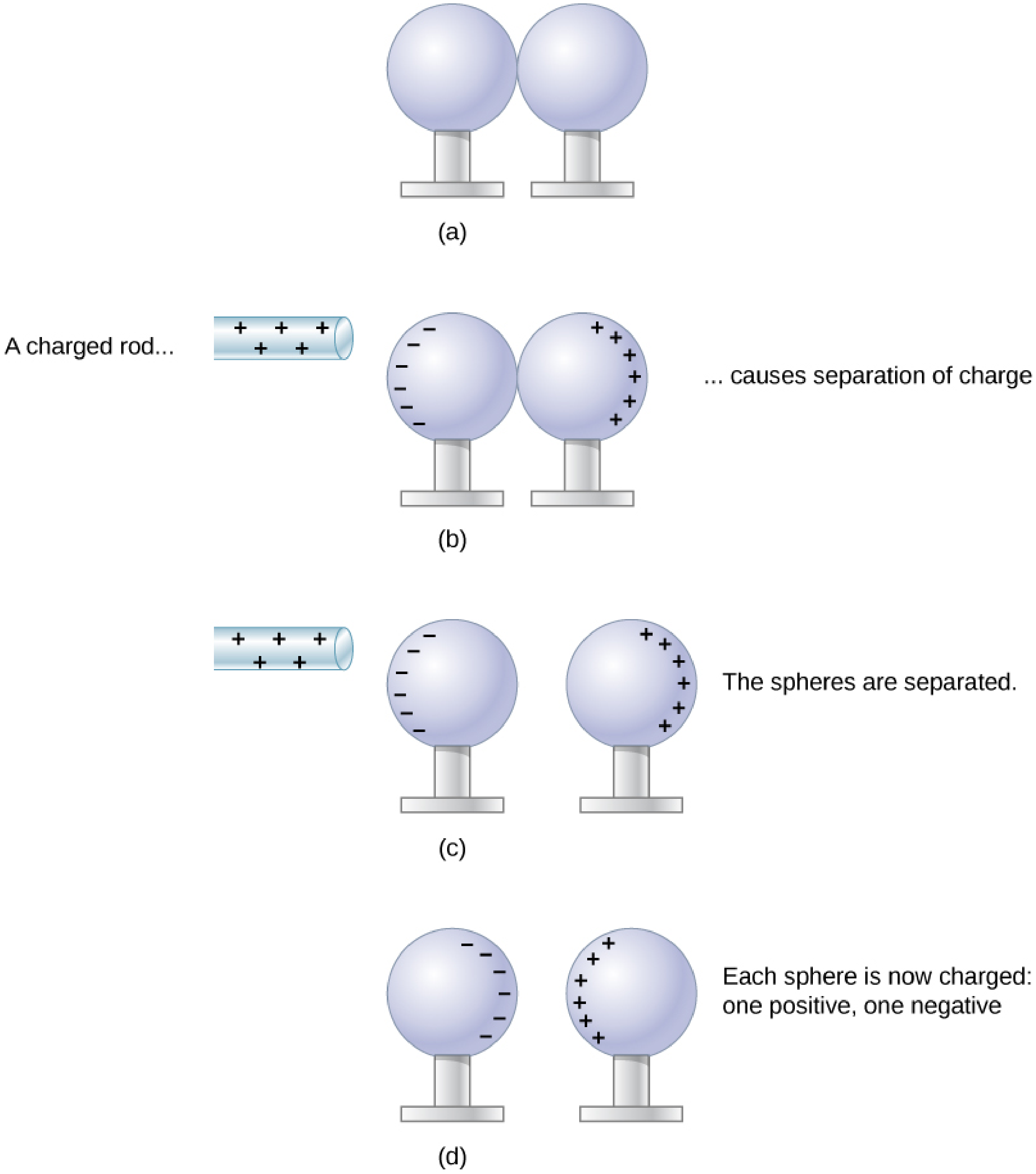
Neutral objects can be attracted to any charged object. The pieces of straw attracted to polished amber are neutral, for example. If you run a plastic comb through your hair, the charged comb can pick up neutral pieces of paper. [link] shows how the polarization of atoms and molecules in neutral objects results in their attraction to a charged object.



Both positive and negative objects attract a neutral object by polarizing its molecules. (a) A positive object brought near a neutral insulator polarizes its molecules. There is a slight shift in the distribution of the electrons orbiting the molecule, with unlike charges being brought nearer and like charges moved away. Since the electrostatic force decreases with distance, there is a net attraction. (b) A negative object produces the opposite polarization, but again attracts the neutral object. (c) The same effect occurs for a conductor; since the unlike charges are closer, there is a net attraction.

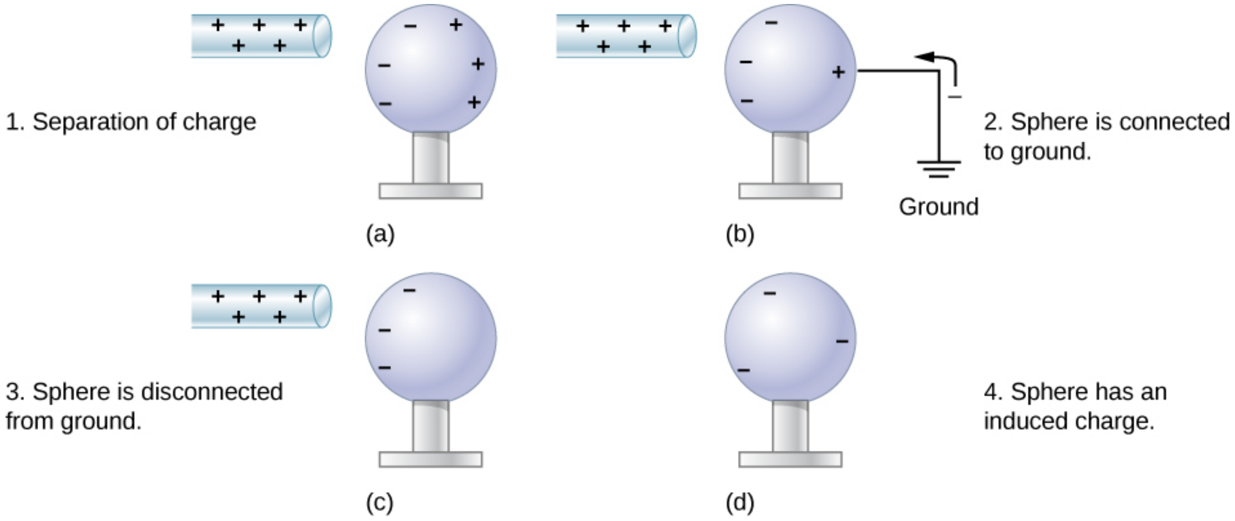
When a charged rod is brought near a neutral substance, an insulator in this case, the distribution of charge in atoms and molecules is shifted slightly. Opposite charge is attracted nearer the external charged rod, while like charge is repelled. Since the electrostatic force decreases with distance, the repulsion of like charges is weaker than the attraction of unlike charges, and so there is a net attraction. Thus, a positively charged glass rod attracts neutral pieces of paper, as will a negatively charged rubber rod. Some molecules, like water, are polar molecules. Polar molecules have a natural or inherent separation of charge, although they are neutral overall. Polar molecules are particularly affected by other charged objects and show greater polarization effects than molecules with naturally uniform charge distributions.

When the two ends of a dipole can be separated, this method of **charging by induction** may be used to create charged objects without transferring charge. In [link], we see two neutral metal spheres in contact with one another but insulated from the rest of the world. A positively charged rod is brought near one of them, attracting negative charge to that side, leaving the other sphere positively charged.


(a)
A charged rod...
... causes separation of charge
(b)
The spheres are separated.
(c)
Each sphere is now charged:
one positive, one negative
(d)

Charging by induction. (a) Two uncharged or neutral metal spheres are in contact with each other but insulated from the rest of the world. (b) A positively charged glass rod is brought near the sphere on the left, attracting negative charge and leaving the other sphere positively charged. (c) The spheres are separated before the rod is removed, thus separating negative and positive charges. (d) The spheres retain net charges after the inducing rod is removed—without ever having been touched by a charged object.

Another method of charging by induction is shown in [link]. The neutral metal sphere is polarized when a charged rod is brought near it. The sphere is then grounded, meaning that a conducting wire is run from the sphere to the ground. Since Earth is large and most of the ground is a good conductor, it can supply or accept excess charge easily. In this case, electrons are attracted to the sphere through a wire called the ground wire, because it supplies a conducting path to the ground. The ground connection is broken before the charged rod is removed, leaving the sphere with an excess charge opposite to that of the rod. Again, an opposite charge is achieved when charging by induction, and the charged rod loses none of its excess charge.


1. Separation of charge
(a)
2. Sphere is connected to ground.
Ground
(b)
3. Sphere is disconnected from ground.
(c)
4. Sphere has an induced charge.
(d)

Charging by induction using a ground connection. (a) A positively charged rod is brought near a neutral metal sphere, polarizing it. (b) The sphere is grounded, allowing electrons to be attracted from Earth’s ample supply. (c) The ground connection is broken. (d) The positive rod is removed, leaving the sphere with an induced negative charge.

Summary

- A conductor is a substance that allows charge to flow freely through its atomic structure.
- An insulator holds charge fixed in place.
- Polarization is the separation of positive and negative charges in a neutral object. Polarized objects have their positive and negative charges concentrated in different

areas, giving them a charge distribution.

Conceptual Questions

Exercise:

Problem:

An eccentric inventor attempts to levitate a cork ball by wrapping it with foil and placing a large negative charge on the ball and then putting a large positive charge on the ceiling of his workshop. Instead, while attempting to place a large negative charge on the ball, the foil flies off. Explain.

Exercise:

Problem:

When a glass rod is rubbed with silk, it becomes positive and the silk becomes negative—yet both attract dust. Does the dust have a third type of charge that is attracted to both positive and negative? Explain.

Solution:

No, the dust is attracted to both because the dust particle molecules become polarized in the direction of the silk.

Exercise:

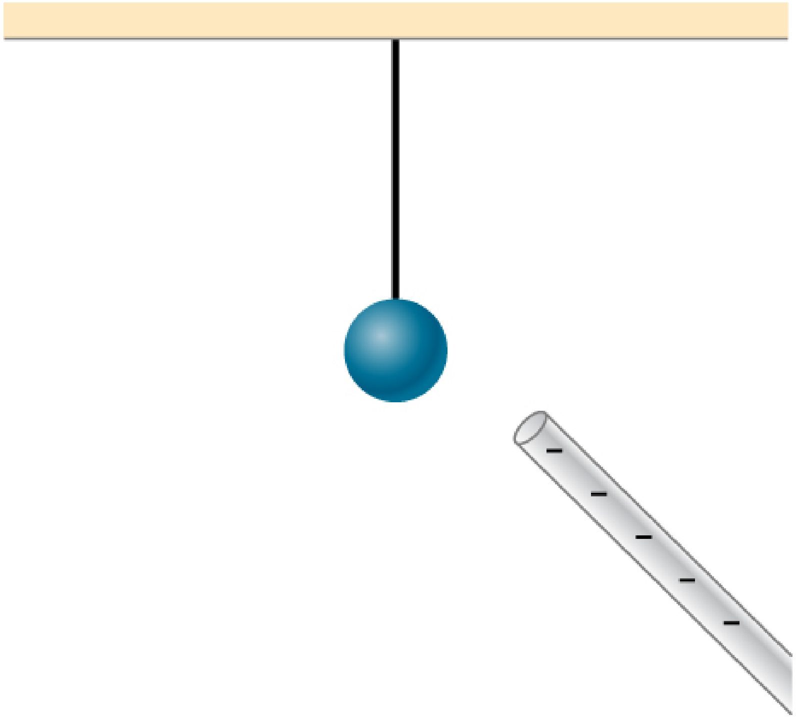
Problem:

Why does a car always attract dust right after it is polished? (Note that car wax and car tires are insulators.)

Exercise:

Problem:

Does the uncharged conductor shown below experience a net electric force?



Solution:

Yes, polarization charge is induced on the conductor so that the positive charge is nearest the charged rod, causing an attractive force.

Exercise:

Problem:

While walking on a rug, a person frequently becomes charged because of the rubbing between his shoes and the rug. This charge then causes a spark and a slight shock when the person gets close to a metal object. Why are these shocks so much more common on a dry day?

Exercise:

Problem: Compare charging by conduction to charging by induction.

Solution:

Charging by conduction is charging by contact where charge is transferred to the object. Charging by induction first involves producing a polarization charge in the object and then connecting a wire to ground to allow some of the charge to leave the object, leaving the object charged.

Exercise:

Problem:

Small pieces of tissue are attracted to a charged comb. Soon after sticking to the comb, the pieces of tissue are repelled from it. Explain.

Exercise:

Problem:

Trucks that carry gasoline often have chains dangling from their undercarriages and brushing the ground. Why?

Solution:

This is so that any excess charge is transferred to the ground, keeping the gasoline receptacles neutral. If there is excess charge on the gasoline receptacle, a spark could ignite it.

Exercise:

Problem: Why do electrostatic experiments work so poorly in humid weather?

Exercise:

Problem:

Why do some clothes cling together after being removed from the clothes dryer? Does this happen if they’re still damp?

Solution:

The dryer charges the clothes. If they are damp, the presence of water molecules suppresses the charge.

Exercise:

Problem: Can induction be used to produce charge on an insulator?

Exercise:

Problem:

Suppose someone tells you that rubbing quartz with cotton cloth produces a third kind of charge on the quartz. Describe what you might do to test this claim.

Solution:

There are only two types of charge, attractive and repulsive. If you bring a charged object near the quartz, only one of these two effects will happen, proving there is not a third kind of charge.

Exercise:

Problem:

A handheld copper rod does not acquire a charge when you rub it with a cloth. Explain why.

Exercise:

Problem:

Suppose you place a charge *q* near a large metal plate. (a) If *q* is attracted to the plate, is the plate necessarily charged? (b) If *q* is repelled by the plate, is the plate necessarily charged?

Solution:

a. No, since a polarization charge is induced. b. Yes, since the polarization charge would produce only an attractive force.

Problems

Exercise:

Problem:

Suppose a speck of dust in an electrostatic precipitator has 1.0000×10^{12} protons in it and has a net charge of −5.00 nC (a very large charge for a small speck). How many electrons does it have?

Solution:

$5.00 \times 10^{-9}\,\text{C}\,(6.242 \times 10^{18}\,\text{e/C}) = 3.121 \times 10^{10}\,\text{e}$;
$3.121 \times 10^{10}\,\text{e} + 1.0000 \times 10^{12}\,\text{e} = 1.0312 \times 10^{12}\,\text{e}$

Exercise:

Problem:

An amoeba has 1.00×10^{16} protons and a net charge of 0.300 pC. (a) How many fewer electrons are there than protons? (b) If you paired them up, what fraction of the protons would have no electrons?

Exercise:

Problem:

A 50.0-g ball of copper has a net charge of $2.00\ \mu\text{C}$. What fraction of the copper's electrons has been removed? (Each copper atom has 29 protons, and copper has an atomic mass of 63.5.)

Solution:

atomic mass of copper atom times $1\text{ u} = 1.055\ \times\ 10^{-25}\text{ kg}$;
number of copper atoms $= 4.739\ \times\ 10^{23}\text{ atoms}$;
number of electrons equals 29 times number of atoms or $1.374\ \times\ 10^{25}\text{ electrons}$;
$\frac{2.00\times10^{-6}\text{ C}(6.242\times10^{18}\text{ e/C})}{1.374\times10^{25}\text{ e}} = 9.083\ \times\ 10^{-13}$ or $9.083\ \times\ 10^{-11}\%$

Exercise:

Problem:

What net charge would you place on a 100-g piece of sulfur if you put an extra electron on 1 in 10^{12} of its atoms? (Sulfur has an atomic mass of 32.1 u.)

Exercise:

Problem:

How many coulombs of positive charge are there in 4.00 kg of plutonium, given its atomic mass is 244 and that each plutonium atom has 94 protons?

Solution:

$244.00\text{ u}(1.66\ \times\ 10^{-27}\text{ kg/u}) = 4.050\ \times\ 10^{-25}\text{ kg}$;
$\frac{4.00\text{ kg}}{4.050\times10^{-25}\text{ kg}} = 9.877\ \times\ 10^{24}\text{ atoms}\quad 9.877\ \times\ 10^{24}(94) = 9.284\ \times\ 10^{26}\text{ protons}$
;
$9.284\ \times\ 10^{26}(1.602\ \times\ 10^{-19}\text{ C/p}) = 1.487\ \times\ 10^{8}\text{ C}$

Glossary

charging by induction	process by which an electrically charged object brought near a neutral object creates a charge separation in that object
conduction electron	electron that is free to move away from its atomic orbit
conductor	

material that allows electrons to move separately from their atomic orbits; object with properties that allow charges to move about freely within it

dipole
two equal and opposite charges that are fixed close to each other

insulator
material that holds electrons securely within their atomic orbits

polarization
slight shifting of positive and negative charges to opposite sides of an object

Coulomb's Law

By the end of this section, you will be able to:

- Describe the electric force, both qualitatively and quantitatively
- Calculate the force that charges exert on each other
- Determine the direction of the electric force for different source charges
- Correctly describe and apply the superposition principle for multiple source charges

Experiments with electric charges have shown that if two objects each have electric charge, then they exert an electric force on each other. The magnitude of the force is linearly proportional to the net charge on each object and inversely proportional to the square of the distance between them. (Interestingly, the force does not depend on the mass of the objects.) The direction of the force vector is along the imaginary line joining the two objects and is dictated by the signs of the charges involved.

Let

- $q_1, q_2 =$ the net electric charges of the two objects;
- $\vec{\mathbf{r}}_{12} =$ the vector displacement from q_1 to q_2.

The electric force $\vec{\mathbf{F}}$ on one of the charges is proportional to the magnitude of its own charge and the magnitude of the other charge, and is inversely proportional to the square of the distance between them:

Equation:

$$F \propto \frac{q_1 q_2}{r_{12}^2}.$$

This proportionality becomes an equality with the introduction of a proportionality constant. For reasons that will become clear in a later chapter, the proportionality constant that we use is actually a collection of constants. (We discuss this constant shortly.)

Note:

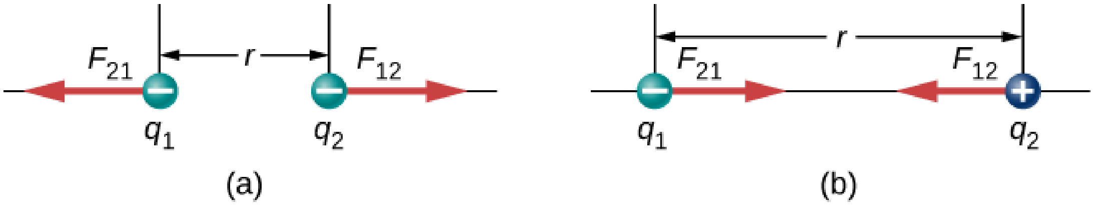
Coulomb’s Law

The magnitude of the electric force (or **Coulomb force**) between two electrically charged particles is equal to

Equation:

$$|\mathbf{F}_{12}| = \frac{1}{4\pi\varepsilon_0} \frac{|q_1 q_2|}{r_{12}^2}$$

The unit vector r has a magnitude of 1 and points along the axis as the charges. If the charges have the same sign, the force is in the same direction as r showing a repelling force. If the charges have different signs, the force is in the opposite direction of r showing an attracting force. ([link]).



The electrostatic force $\vec{\mathbf{F}}$ between point charges q_1 and q_2 separated by a distance r is given by Coulomb's law. Note that Newton's third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges; (b) unlike charges.

It is important to note that the electric force is not constant; it is a function of the separation distance between the two charges. If either the test charge or the source charge (or both) move, then $\vec{\mathbf{r}}$ changes, and therefore so does the force. An immediate consequence of this is that direct application of Newton's laws with this force can be mathematically difficult, depending on the specific problem at hand. It can (usually) be done, but we almost always look for easier methods of calculating whatever physical quantity we are interested in. (Conservation of energy is the most common choice.)

Finally, the new constant ε_0 in Coulomb's law is called the *permittivity of free space*, or (better) the **permittivity of vacuum**. It has a very important physical meaning that we will discuss in a later chapter; for now, it is simply an empirical proportionality constant. Its numerical value (to three significant figures) turns out to be

Equation:

$$\varepsilon_0 = 8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N} \cdot \mathrm{m}^2}.$$

These units are required to give the force in Coulomb's law the correct units of newtons. Note that in Coulomb's law, the permittivity of vacuum is only part of the proportionality constant. For convenience, we often define a Coulomb's constant:

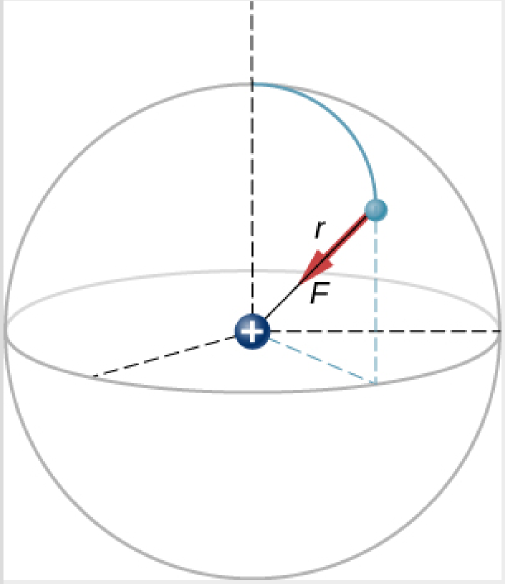
Equation:

$$k_e = \frac{1}{4\pi\varepsilon_0} = 8.99 \times 10^9 \frac{\mathrm{N} \cdot \mathrm{m}^2}{\mathrm{C}^2}.$$

Example:

The Force on the Electron in Hydrogen

A hydrogen atom consists of a single proton and a single electron. The proton has a charge of $+e$ and the electron has $-e$. In the “ground state” of the atom, the electron orbits the proton at most probable distance of 5.29×10^{-11} m ([link]). Calculate the electric force on the electron due to the proton.



A schematic depiction of a hydrogen atom, showing the force on the electron. This depiction is only to enable us to calculate the force; the hydrogen atom does not really look like this. Recall [link].

Strategy

For the purposes of this example, we are treating the electron and proton as two point particles, each with an electric charge, and we are told the distance between them; we are asked to calculate the force on the electron. We thus use Coulomb’s law.

Solution

Our two charges and the distance between them are,

Equation:

$$\begin{aligned} q_1 &= +e = +1.602 \times 10^{-19}\,\text{C} \\ q_2 &= -e = -1.602 \times 10^{-19}\,\text{C} \\ r &= 5.29 \times 10^{-11}\,\text{m}. \end{aligned}$$

The magnitude of the force on the electron is

Equation:

$$F = \frac{1}{4\pi\epsilon_0}\frac{|e|^2}{r^2} = \frac{1}{4\pi\left(8.85\times 10^{-12}\,\frac{\mathrm{C}^2}{\mathrm{N\cdot m^2}}\right)}\frac{\left(1.602\times 10^{-19}\,\mathrm{C}\right)^2}{\left(5.29\times 10^{-11}\,\mathrm{m}\right)^2} = 8.25\times 10^{-8}\,\mathrm{N}.$$

As for the direction, since the charges on the two particles are opposite, the force is attractive; the force on the electron points radially directly toward the proton, everywhere in the electron’s orbit. The force is thus expressed as

Equation:

$$\vec{\mathbf{F}} = \left(8.25\times 10^{-8}\,\mathrm{N}\right)\hat{\mathbf{r}}.$$

Significance

This is a three-dimensional system, so the electron (and therefore the force on it) can be anywhere in an imaginary spherical shell around the proton. In this “classical” model of the hydrogen atom, the electrostatic force on the electron points in the inward centripetal direction, thus maintaining the electron’s orbit. But note that the quantum mechanical model of hydrogen (discussed in Quantum Mechanics) is utterly different.

Note:

Exercise:

Problem:

Check Your Understanding What would be different if the electron also had a positive charge?

Solution:

The force would point outward.

Multiple Source Charges

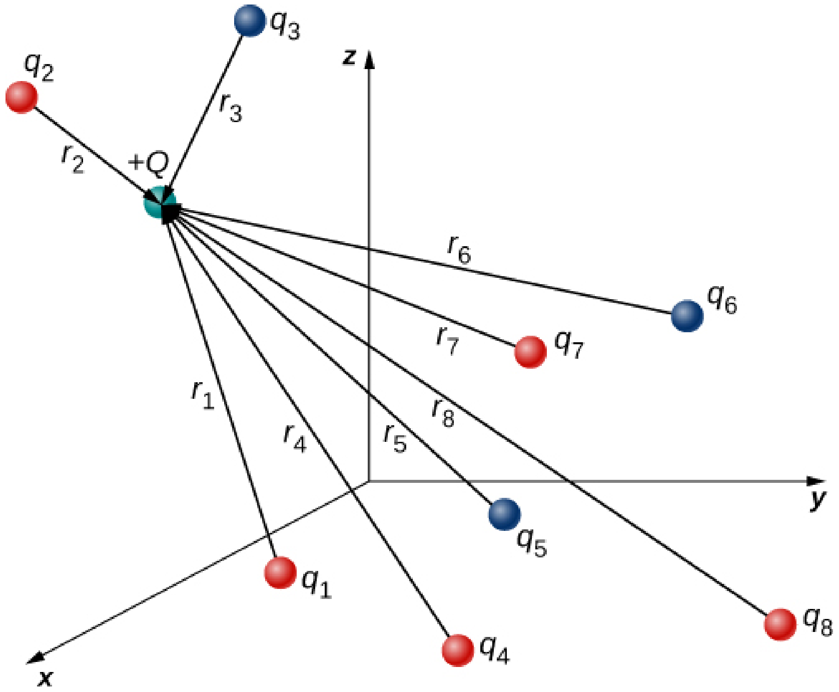
The analysis that we have done for two particles can be extended to an arbitrary number of particles; we simply repeat the analysis, two charges at a time. Specifically, we ask the question: Given N charges (which we refer to as source charge), what is the net electric force that they exert on some other point charge (which we call the test charge)? Note that we use these terms because we can think of the test charge being used to test the strength of the force provided by the source charges.

Like all forces that we have seen up to now, the net electric force on our test charge is simply the vector sum of each individual electric force exerted on it by each of the individual test charges. Thus, we can calculate the net force on the test charge Q by calculating the force on it from each source charge, taken one at a time, and then adding all those forces together (as vectors). This ability to simply add up individual forces in this way is referred to as the **principle of superposition**, and is one of the more important features of the electric force. In mathematical form, this becomes

Note:
Equation:

$$\vec{\mathbf{F}}(r) = \frac{1}{4\pi\varepsilon_0} Q \sum_{i=1}^{N} \frac{q_i}{r_i^2} \hat{\mathbf{r}}_i.$$

In this expression, Q represents the charge of the particle that is experiencing the electric force $\vec{\mathbf{F}}$, and is located at $\vec{\mathbf{r}}$ from the origin; the q_i's are the N source charges, and the vectors $\vec{\mathbf{r}}_i = r_i \hat{\mathbf{r}}_i$ are the displacements from the position of the ith charge to the position of Q. Each of the N unit vectors points directly from its associated source charge toward the test charge. All of this is depicted in [link]. Please note that there is no physical difference between Q and q_i; the difference in labels is merely to allow clear discussion, with Q being the charge we are determining the force on.



The eight source charges each apply a force on the single test charge Q. Each force can be calculated independently of the other seven forces. This is the essence of the superposition principle.

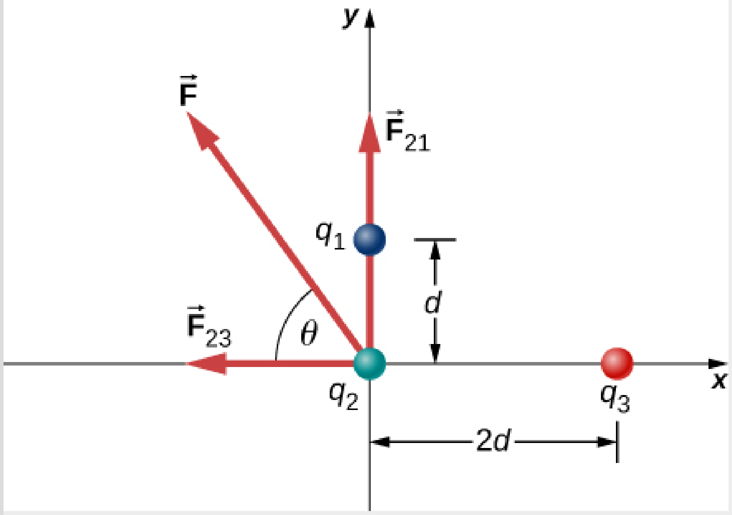
(Note that the force vector $\vec{\mathbf{F}}_i$ does not necessarily point in the same direction as the unit vector $\hat{\mathbf{r}}_i$; it may point in the opposite direction, $-\hat{\mathbf{r}}_i$. The signs of the source charge and test charge determine the direction of the force on the test charge.)

There is a complication, however. Just as the source charges each exert a force on the test charge, so too (by Newton's third law) does the test charge exert an equal and opposite force on each of the source charges. As a consequence, each source charge would change position. However, by [link], the force on the test charge is a function of position; thus, as the positions of the source charges change, the net force on the test charge necessarily changes, which changes the force, which again changes the positions. Thus, the entire mathematical analysis quickly becomes intractable. Later, we will learn techniques for handling this situation, but for now, we make the simplifying assumption that the source charges are fixed in place somehow, so that their positions are constant in time. (The test charge is allowed to move.) With this restriction in place, the analysis of charges is known as **electrostatics**, where "statics" refers to the constant (that is, static) positions of the source charges and the force is referred to as an **electrostatic force**.

Example:

The Net Force from Two Source Charges

Three different, small charged objects are placed as shown in [link]. The charges q_1 and q_3 are fixed in place; q_2 is free to move. Given $q_1 = 2e$, $q_2 = -3e$, and $q_3 = -5e$, and that $d = 2.0 \times 10^{-7}$ m, what is the net force on the middle charge q_2?



Source charges q_1 and q_3 each apply a force on q_2.

Strategy

We use Coulomb's law again. The way the question is phrased indicates that q_2 is our test charge, so that q_1 and q_3 are source charges. The principle of superposition says that the force on q_2 from each of the other charges is unaffected by the presence of the other charge. Therefore, we write down the force on q_2 from each and add them together as vectors.

Solution

We have two source charges $(q_1$ and $q_3)$, a test charge (q_2), distances $(r_{21}$ and $r_{23})$, and we are asked to find a force. This calls for Coulomb's law and superposition of forces. There are two forces:

Equation:

$$\vec{\mathbf{F}} = \vec{\mathbf{F}}_{21} + \vec{\mathbf{F}}_{23} = \frac{1}{4\pi\varepsilon_0}\left[\frac{q_2 q_1}{r_{21}^2}\hat{\mathbf{j}} + \left(-\frac{q_2 q_3}{r_{23}^2}\hat{\mathbf{i}}\right)\right].$$

We can't add these forces directly because they don't point in the same direction: $\vec{\mathbf{F}}_{12}$ points only in the −*x*-direction, while $\vec{\mathbf{F}}_{13}$ points only in the +*y*-direction. The net force is obtained from applying the Pythagorean theorem to its *x*- and *y*-components:

Equation:

$$F=\sqrt{F_x^2+F_y^2}$$

where

Equation:

$$\begin{aligned} F_x &= -F_{23} = -\frac{1}{4\pi\varepsilon_0}\,\frac{q_2q_3}{r_{23}^2} \\ &= -\left(8.99\times10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{(4.806\times10^{-19}\,\text{C})(8.01\times10^{-19}\,\text{C})}{(4.00\times10^{-7}\,\text{m})^2} \\ &= -2.16\times10^{-14}\,\text{N} \end{aligned}$$

and

Equation:

$$\begin{aligned} F_y &= F_{21} = \frac{1}{4\pi\varepsilon_0}\,\frac{q_2q_1}{r_{21}^2} \\ &= \left(8.99\times10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{(4.806\times10^{-19}\,\text{C})(3.204\times10^{-19}\,\text{C})}{(2.00\times10^{-7}\,\text{m})^2} \\ &= 3.46\times10^{-14}\,\text{N}. \end{aligned}$$

We find that

Equation:

$$F=\sqrt{F_x^2+F_y^2}=4.08\times10^{-14}\,\text{N}$$

at an angle of

Equation:

$$\phi=\tan^{-1}\left(\frac{F_y}{F_x}\right)=\tan^{-1}\left(\frac{3.46\times10^{-14}\,\text{N}}{-2.16\times10^{-14}\,\text{N}}\right)=-58^\circ,$$

that is, 58° above the −*x*-axis, as shown in the diagram.

Significance

Notice that when we substituted the numerical values of the charges, we did not include the negative sign of either q_2 or q_3. Recall that negative signs on vector quantities indicate a reversal of direction of the vector in question. But for electric forces, the direction of the force is determined by the types (signs) of both interacting charges; we determine the force directions by considering whether the signs of the two charges are the same or are opposite. If you also include negative signs from negative charges when you substitute numbers, you run the risk of mathematically reversing the direction of the force you are calculating. Thus, the safest thing to do is to calculate just the magnitude of the force, using the absolute values of the charges, and determine the directions physically.

It’s also worth noting that the only new concept in this example is how to calculate the electric forces; everything else (getting the net force from its components, breaking the forces into their components, finding the direction of the net force) is the same as force problems you have done earlier.

Note:
Exercise:

Problem: **Check Your Understanding** What would be different if q_1 were negative?

Solution:

The net force would point 58° below the −*x*-axis.

Summary

- Coulomb's law gives the magnitude of the force vector between point charges. It is **Equation:**

$$\vec{\mathbf{F}}_{12}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{\mathbf{r}}_{12}$$

  where q_1 and q_2 are two point charges separated by a distance *r*. This Coulomb force is extremely basic, since most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.

Conceptual Questions

Exercise:

Problem:

Would defining the charge on an electron to be positive have any effect on Coulomb's law?

Exercise:

Problem:

An atomic nucleus contains positively charged protons and uncharged neutrons. Since nuclei do stay together, what must we conclude about the forces between these nuclear particles?

Solution:

The force holding the nucleus together must be greater than the electrostatic repulsive force on the protons.

Exercise:

Problem:

Is the force between two fixed charges influenced by the presence of other charges?

Problems

Exercise:

Problem:

Two point particles with charges $+3\ \mu\text{C}$ and $+5\ \mu\text{C}$ are held in place by 3-N forces on each charge in appropriate directions. (a) Draw a free-body diagram for each particle. (b) Find the distance between the charges.

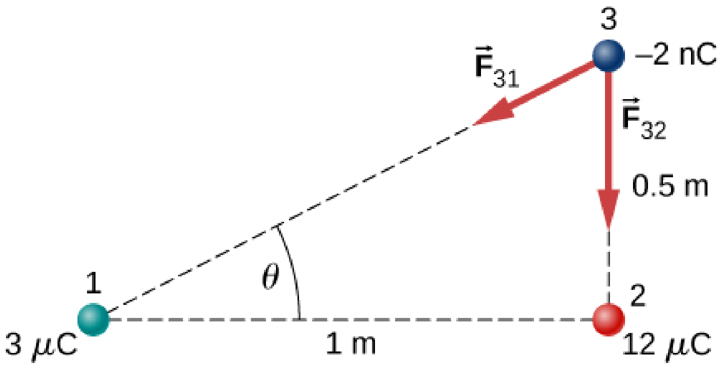
Exercise:

Problem:

Two charges $+3\ \mu\text{C}$ and $+12\ \mu\text{C}$ are fixed 1 m apart, with the second one to the right. Find the magnitude and direction of the net force on a −2-nC charge when placed at the following locations: (a) halfway between the two (b) half a meter to the left of the $+3\ \mu\text{C}$ charge (c) half a meter above the $+12\ \mu\text{C}$ charge in a direction perpendicular to the line joining the two fixed charges

Solution:

a. charge 1 is $3\ \mu\text{C}$; charge 2 is $12\ \mu\text{C}$, $\text{F}_{31} = 2.16\ \times\ 10^{-4}$ N to the left,
$\text{F}_{32} = 8.63\ \times\ 10^{-4}$ N to the right,
$\text{F}_{\text{net}} = 6.47\ \times\ 10^{-4}$ N to the right;
b. $\text{F}_{31} = 2.16\ \times\ 10^{-4}$ N to the right,
$\text{F}_{32} = 9.59\ \times\ 10^{-5}$ N to the right,
$\text{F}_{\text{net}} = 3.12\ \times\ 10^{-4}$ N to the right,



;

c. $\vec{\mathbf{F}}_{31x} = -2.76 \times 10^{-5}\,\text{N}\,\hat{\mathbf{i}}$,
$\vec{\mathbf{F}}_{31y} = -1.38 \times 10^{-5}\,\text{N}\,\hat{\mathbf{j}}$,
$\vec{\mathbf{F}}_{32y} = -8.63 \times 10^{-4}\,\text{N}\,\hat{\mathbf{j}}$
$\vec{\mathbf{F}}_{\text{net}} = -3.86 \times 10^{-5}\,\text{N}\,\hat{\mathbf{i}} - 8.83 \times 10^{-4}\,\text{N}\,\hat{\mathbf{j}}$

Exercise:

Problem:

In a salt crystal, the distance between adjacent sodium and chloride ions is 2.82×10^{-10} m. What is the force of attraction between the two singly charged ions?

Exercise:

Problem:

Protons in an atomic nucleus are typically 10^{-15} m apart. What is the electric force of repulsion between nuclear protons?

Solution:

$F = 230.7\,\text{N}$

Exercise:

Problem:

Suppose Earth and the Moon each carried a net negative charge $-Q$. Approximate both bodies as point masses and point charges.

(a) What value of Q is required to balance the gravitational attraction between Earth and the Moon?

(b) Does the distance between Earth and the Moon affect your answer? Explain.

(c) How many electrons would be needed to produce this charge?

Exercise:

Problem:

Point charges $q_1 = 50\ \mu\text{C}$ and $q_2 = -25\ \mu\text{C}$ are placed 1.0 m apart. What is the force on a third charge $q_3 = 20\ \mu\text{C}$ placed midway between q_1 and q_2?

Solution:

$F = 53.94\ \text{N}$

Exercise:

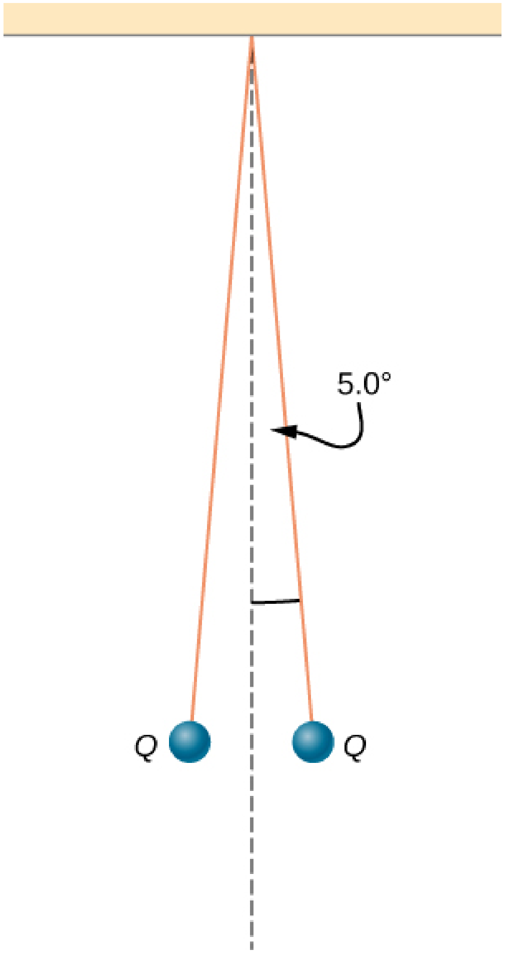
Problem:

Where must q_3 of the preceding problem be placed so that the net force on it is zero?

Exercise:

Problem:

Two small balls, each of mass 5.0 g, are attached to silk threads 50 cm long, which are in turn tied to the same point on the ceiling, as shown below. When the balls are given the same charge *Q*, the threads hang at 5.0° to the vertical, as shown below. What is the magnitude of *Q*? What are the signs of the two charges?



Solution:

The tension is $T=0.049\,\text{N}$. The horizontal component of the tension is $0.0043\,\text{N}$
$d=0.088\,\text{m},\quad q=6.1\ \times\ 10^{-8}\ \text{C}$.
The charges can be positive or negative, but both have to be the same sign.

Exercise:

Problem:

Point charges $Q_1=2.0\ \mu\text{C}$ and $Q_2=4.0\ \mu\text{C}$ are located at $\vec{\mathbf{r}}_1=(4.0\hat{\mathbf{i}}-2.0\hat{\mathbf{j}}+5.0\widehat{\mathbf{k}})\text{m}$ and $\vec{\mathbf{r}}_2=(8.0\hat{\mathbf{i}}+5.0\hat{\mathbf{j}}-9.0\widehat{\mathbf{k}})\text{m}$. What is the force of Q_2 on Q_1?

Exercise:

Problem:

The net excess charge on two small spheres (small enough to be treated as point charges) is Q. Show that the force of repulsion between the spheres is greatest when each sphere has an excess charge $Q/2$. Assume that the distance between the spheres is so large compared with their radii that the spheres can be treated as point charges.

Solution:

Let the charge on one of the spheres be nQ, where n is a fraction between 0 and 1. In the numerator of Coulomb's law, the term involving the charges is $nQ(1-n)Q$. This is equal to $(n-n^2)Q^2$. Finding the maximum of this term gives $1-2n=0 \Rightarrow n=\frac{1}{2}$

Exercise:

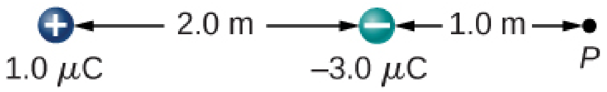
Problem:

Two small, identical conducting spheres repel each other with a force of 0.050 N when they are 0.25 m apart. After a conducting wire is connected between the spheres and then removed, they repel each other with a force of 0.060 N. What is the original charge on each sphere?

Exercise:

Problem:

A charge $q=2.0\ \mu\text{C}$ is placed at the point P shown below. What is the force on q?



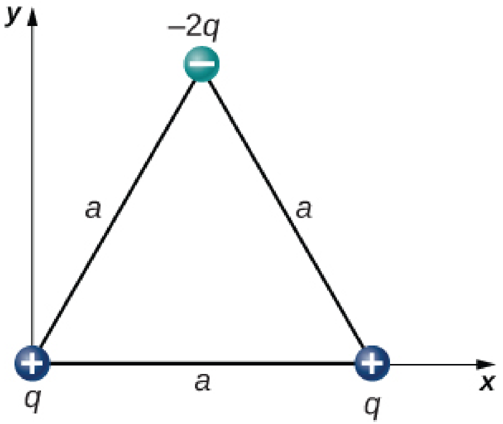
Solution:

Define right to be the positive direction and hence left is the negative direction, then $F=-0.05\ \text{N}$

Exercise:

Problem:

What is the net electric force on the charge located at the lower right-hand corner of the triangle shown here?



Exercise:

Problem:

Two fixed particles, each of charge $5.0\ \times\ 10^{-6}$ C, are 24 cm apart. What force do they exert on a third particle of charge $-2.5\ \times\ 10^{-6}$ C that is 13 cm from each of them?

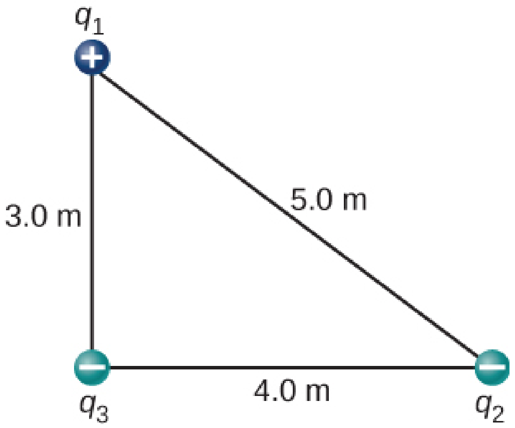
Solution:

The particles form triangle of sides 13, 13, and 24 cm. The *x*-components cancel, whereas there is a contribution to the *y*-component from both charges 24 cm apart. The *y*-axis passing through the third charge bisects the 24-cm line, creating two right triangles of sides 5, 12, and 13 cm.
$F_y = 2.56$ N in the negative *y*-direction since the force is attractive. The net force from both charges is $\vec{\mathbf{F}}_{\text{net}} = -5.12\ \text{N}\hat{\mathbf{j}}$.

Exercise:

Problem:

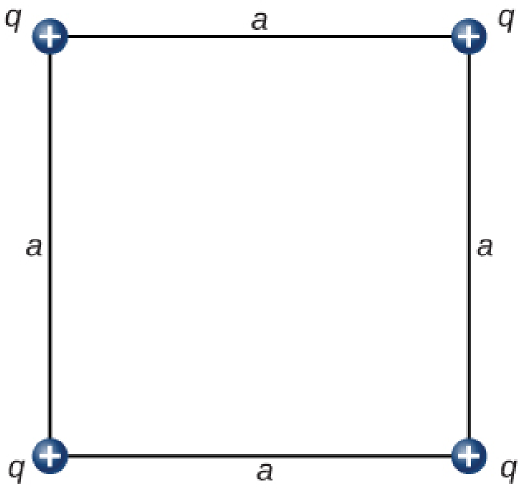
The charges $q_1 = 2.0\ \times\ 10^{-7}$ C, $q_2 = -4.0\ \times\ 10^{-7}$ C, and $q_3 = -1.0\ \times\ 10^{-7}$ C are placed at the corners of the triangle shown below. What is the force on q_1?



Exercise:

Problem:

What is the force on the charge q at the lower-right-hand corner of the square shown here?



Solution:

The diagonal is $\sqrt{2}a$ and the components of the force due to the diagonal charge has a factor $\cos\theta = \frac{1}{\sqrt{2}}$;

$$\vec{\mathbf{F}}_{\text{net}} = \left[k\frac{q^2}{a^2} + k\frac{q^2}{2a^2}\ \frac{1}{\sqrt{2}}\right]\hat{\mathbf{i}} - \left[k\frac{q^2}{a^2} + k\frac{q^2}{2a^2}\ \frac{1}{\sqrt{2}}\right]\hat{\mathbf{j}}$$

Exercise:

Problem:

Point charges $q_1 = 10\ \mu\text{C}$ and $q_2 = -30\ \mu\text{C}$ are fixed at $r_1 = \left(3.0\hat{\mathbf{i}} - 4.0\hat{\mathbf{j}}\right)\text{m}$ and $r_2 = \left(9.0\hat{\mathbf{i}} + 6.0\hat{\mathbf{j}}\right)\text{m}$. What is the force of q_2 on q_1?

Glossary

Coulomb force
another term for the electrostatic force

Coulomb’s law
mathematical equation calculating the electrostatic force vector between two charged particles

electrostatic force
amount and direction of attraction or repulsion between two charged bodies; the assumption is that the source charges have no acceleration

electrostatics
study of charged objects which are not in motion

permittivity of vacuum
also called the permittivity of free space, and constant describing the strength of the electric force in a vacuum

principle of superposition
useful fact that we can simply add up all of the forces due to charges acting on an object

Electric Field

By the end of this section, you will be able to:

- Explain the purpose of the electric field concept
- Describe the properties of the electric field
- Calculate the field of a collection of source charges of either sign

As we showed in the preceding section, the net electric force on a test charge is the vector sum of all the electric forces acting on it, from all of the various source charges, located at their various positions. But what if we use a different test charge, one with a different magnitude, or sign, or both? Or suppose we have a dozen different test charges we wish to try at the same location? We would have to calculate the sum of the forces from scratch. Fortunately, it is possible to define a quantity, called the **electric field**, which is independent of the test charge. It only depends on the configuration of the source charges, and once found, allows us to calculate the force on any test charge.

Defining a Field

Suppose we have *N* source charges $q_1, q_2, q_3, \ldots, q_N$ located at positions $\vec{\mathbf{r}}_1, \vec{\mathbf{r}}_2, \vec{\mathbf{r}}_3, \ldots, \vec{\mathbf{r}}_N$, applying *N* electrostatic forces on a test charge Q. The net force on Q is (see [link])

Equation:

$$\begin{aligned}
\vec{\mathbf{F}} &= \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 + \vec{\mathbf{F}}_3 + \cdots + \vec{\mathbf{F}}_N \\
&= \frac{1}{4\pi\varepsilon_0}\left(\frac{Qq_1}{r_1^2}\hat{\mathbf{r}}_1 + \frac{Qq_2}{r_2^2}\hat{\mathbf{r}}_2 + \frac{Qq_3}{r_3^2}\hat{\mathbf{r}}_3 + \cdots + \frac{Qq_N}{r_1^2}\hat{\mathbf{r}}_N\right) \\
&= Q\left[\frac{1}{4\pi\varepsilon_0}\left(\frac{q_1}{r_1^2}\hat{\mathbf{r}}_1 + \frac{q_2}{r_2^2}\hat{\mathbf{r}}_2 + \frac{q_3}{r_3^2}\hat{\mathbf{r}}_3 + \cdots + \frac{q_N}{r_1^2}\hat{\mathbf{r}}_N\right)\right].
\end{aligned}$$

We can rewrite this as

Note:
Equation:

$$\vec{\mathbf{F}} = Q\vec{\mathbf{E}}$$

where

Equation:

$$\vec{\mathbf{E}} \equiv \frac{1}{4\pi\varepsilon_0}\left(\frac{q_1}{r_1^2}\hat{\mathbf{r}}_1 + \frac{q_2}{r_2^2}\hat{\mathbf{r}}_2 + \frac{q_3}{r_3^2}\hat{\mathbf{r}}_3 + \cdots + \frac{q_N}{r_1^2}\hat{\mathbf{r}}_N\right)$$

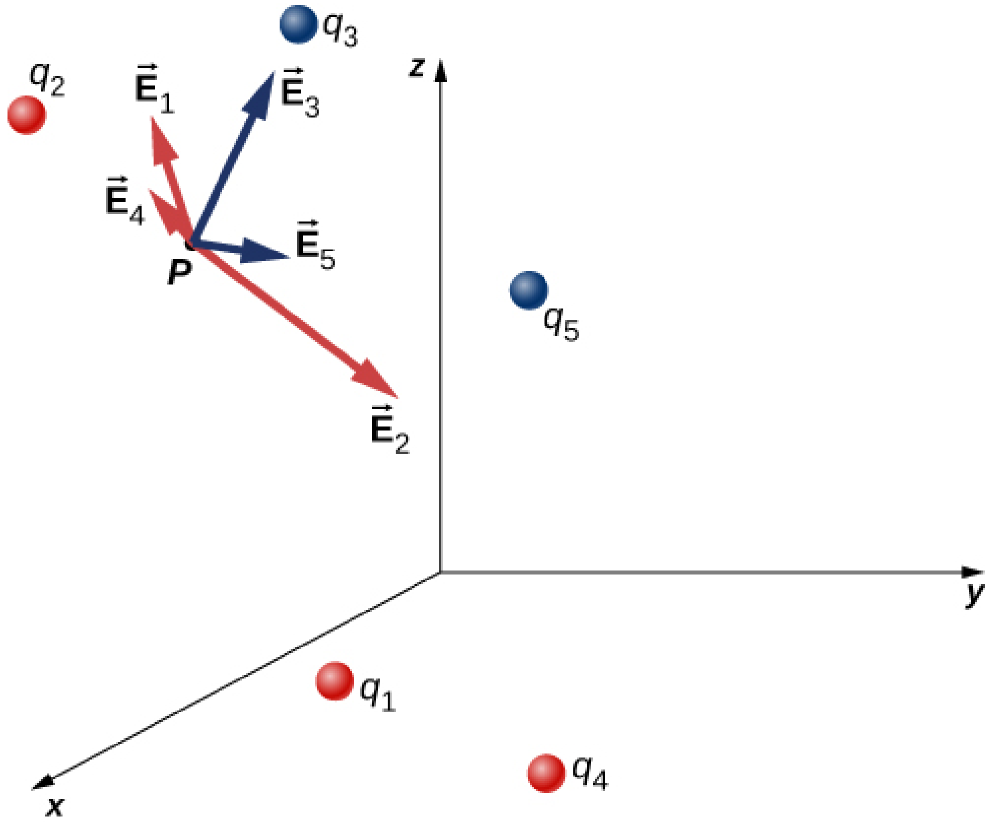
or, more compactly,

Note:

Equation:

$$\vec{\mathbf{E}}(P) \equiv \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{\mathbf{r}}_i.$$

This expression is called the electric field at position $P = P(x, y, z)$ of the N source charges. Here, P is the location of the point in space where you are calculating the field and is relative to the positions $\vec{\mathbf{r}}_i$ of the source charges ([link]). Note that we have to impose a coordinate system to solve actual problems.



Each of these eight source charges creates its own electric field at every point in space; shown here are the field vectors at an arbitrary point *P*. Like the electric force, the net electric field obeys the superposition principle.

Notice that the calculation of the electric field makes no reference to the test charge. Thus, the physically useful approach is to calculate the electric field and then use it to calculate the force on some test charge later, if needed. Different test charges experience different forces [link], but it is the same electric field [link]. That being said, recall that there is no fundamental difference between a test charge and a source charge; these are merely convenient labels for the system of interest. Any charge produces an electric field; however, just as Earth’s orbit is not affected by Earth’s own gravity, a charge is not subject to a force due to the electric field it generates. Charges are only subject to forces from the electric fields of other charges.

In this respect, the electric field $\vec{\mathbf{E}}$ of a point charge is similar to the gravitational field $\vec{\mathbf{g}}$ of Earth; once we have calculated the gravitational field at some point in space, we can use it any time we want to calculate the resulting force on any mass we choose to place at that point. In fact, this is exactly what we do when we say the gravitational field of Earth (near Earth's surface) has a value of $9.81\ \text{m/s}^2$, and then we calculate the resulting force (i.e., weight) on different masses. Also, the general expression for calculating $\vec{\mathbf{g}}$ at arbitrary distances from the center of Earth (i.e., not just near Earth's surface) is very similar to the expression for $\vec{\mathbf{E}}$: $\vec{\mathbf{g}} = G\frac{M}{r^2}\hat{\mathbf{r}}$, where G is a proportionality constant, playing the same role for $\vec{\mathbf{g}}$ as $\frac{1}{4\pi\varepsilon_0}$ does for $\vec{\mathbf{E}}$. The value of $\vec{\mathbf{g}}$ is calculated once and is then used in an endless number of problems.

To push the analogy further, notice the units of the electric field: From $F = QE$, the units of E are newtons per coulomb, N/C, that is, the electric field applies a force on each unit charge. Now notice the units of g: From $w = mg$, the units of g are newtons per kilogram, N/kg, that is, the gravitational field applies a force on each unit mass. We could say that the gravitational field of Earth, near Earth's surface, has a value of 9.81 N/kg.

The Meaning of "Field"

Recall from your studies of gravity that the word "field" in this context has a precise meaning. A field, in physics, is a physical quantity whose value depends on (is a function of) position, relative to the source of the field. In the case of the electric field, [link] shows that the value of $\vec{\mathbf{E}}$ (both the magnitude and the direction) depends on where in space the point P is located, measured from the locations $\vec{\mathbf{r}}_i$ of the source charges q_i.

In addition, since the electric field is a vector quantity, the electric field is referred to as a *vector field*. (The gravitational field is also a vector field.) In contrast, a field that has only a magnitude at every point is a *scalar field*. The temperature in a room is an example of a scalar field. It is a field because the temperature, in general, is different at different locations in the room, and it is a scalar field because temperature is a scalar quantity.

Also, as you did with the gravitational field of an object with mass, you should picture the electric field of a charge-bearing object (the source charge) as a continuous, immaterial substance that surrounds the source charge, filling all of space—in principle, to $\pm\infty$ in all directions. The field exists at every physical point in space. To put it another way, the electric charge on an object alters the space around the charged object in such a way that all other electrically charged objects in space experience an electric force as a result of being in that field. The electric field, then, is the mechanism by which the electric properties of the source charge are transmitted to and through the rest of the universe. (Again, the range of the electric force is infinite.)

We will see in subsequent chapters that the speed at which electrical phenomena travel is the same as the speed of light. There is a deep connection between the electric field and light.

Superposition

Yet another experimental fact about the field is that it obeys the superposition principle. In this context, that means that we can (in principle) calculate the total electric field of many source charges by calculating the electric field of only q_1 at position P, then calculate the field of q_2 at P, while—and this is the crucial idea—ignoring the field of, and indeed even the existence of, q_1. We can repeat this process, calculating the field of each individual source charge, independently of the existence of any of the other charges. The total electric field, then, is the vector sum of all these fields. That, in essence, is what [link] says.

In the next section, we describe how to determine the shape of an electric field of a source charge distribution and how to sketch it.

The Direction of the Field

[link] enables us to determine the magnitude of the electric field, but we need the direction also. We use the convention that the direction of any electric field vector is the same as the direction of the electric force vector that the field would apply to a positive test charge placed in that field. Such a charge would be repelled by positive source charges (the force on it would point

away from the positive source charge) but attracted to negative charges (the force points toward the negative source).

Note:
Direction of the Electric Field

By convention, all electric fields $\vec{\mathbf{E}}$ point away from positive source charges and point toward negative source charges.

Note:
Add charges to the Electric Field of Dreams and see how they react to the electric field. Turn on a background electric field and adjust the direction and magnitude.

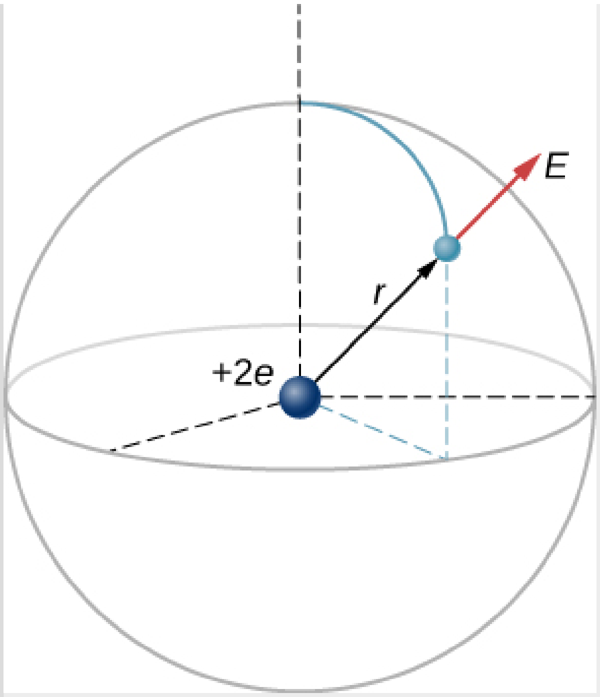
Example:

The *E*-field of an Atom

In an ionized helium atom, the most probable distance between the nucleus and the electron is $r = 26.5 \times 10^{-12}$ m. What is the electric field due to the nucleus at the location of the electron?

Strategy

Note that although the electron is mentioned, it is not used in any calculation. The problem asks for an electric field, not a force; hence, there is only one charge involved, and the problem specifically asks for the field due to the nucleus. Thus, the electron is a red herring; only its distance matters. Also, since the distance between the two protons in the nucleus is much, much smaller than the distance of the electron from the nucleus, we can treat the two protons as a single charge +2*e* ([link]).



A schematic representation of a helium atom. Again, helium physically looks nothing like this, but this sort of diagram is helpful for calculating the electric field of the nucleus.

Solution
The electric field is calculated by
Equation:

$$\vec{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{q_i}{r_i^2} \hat{\mathbf{r}}_i.$$

Since there is only one source charge (the nucleus), this expression simplifies to
Equation:

$$\vec{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0} \frac{q}{r^2} \hat{r}.$$

Here $q = 2e = 2\left(1.6\ \times\ 10^{-19}\ \text{C}\right)$ (since there are two protons) and r is given; substituting gives

Equation:

$$\vec{\mathbf{E}} = \frac{1}{4\pi\left(8.85\ \times\ 10^{-12}\frac{\text{C}^2}{\text{N}\cdot\text{m}^2}\right)}\ \frac{2\left(1.6\ \times\ 10^{-19}\ \text{C}\right)}{\left(26.5\ \times\ 10^{-12}\ \text{m}\right)^2}\hat{\mathbf{r}} = 4.1\ \times\ 10^{12}\frac{\text{N}}{\text{C}}\hat{\mathbf{r}}.$$

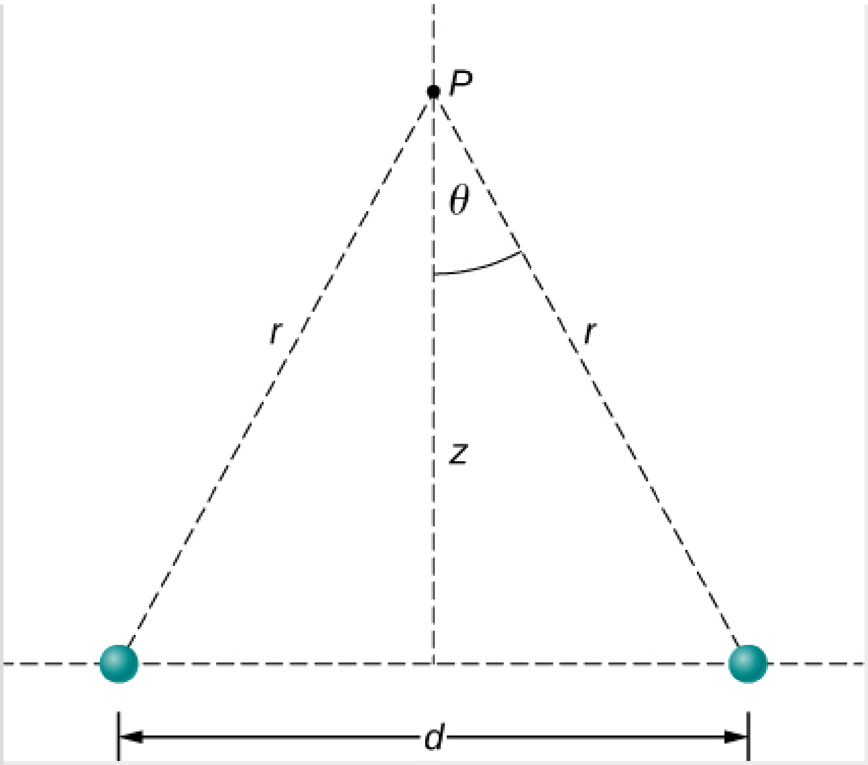
The direction of $\vec{\mathbf{E}}$ is radially away from the nucleus in all directions. Why? Because a positive test charge placed in this field would accelerate radially away from the nucleus (since it is also positively charged), and again, the convention is that the direction of the electric field vector is defined in terms of the direction of the force it would apply to positive test charges.

Example:

The *E*-Field above Two Equal Charges

(a) Find the electric field (magnitude and direction) a distance z above the midpoint between two equal charges $+q$ that are a distance d apart ([link]). Check that your result is consistent with what you'd expect when $z \gg d$.

(b) The same as part (a), only this time make the right-hand charge $-q$ instead of $+q$.

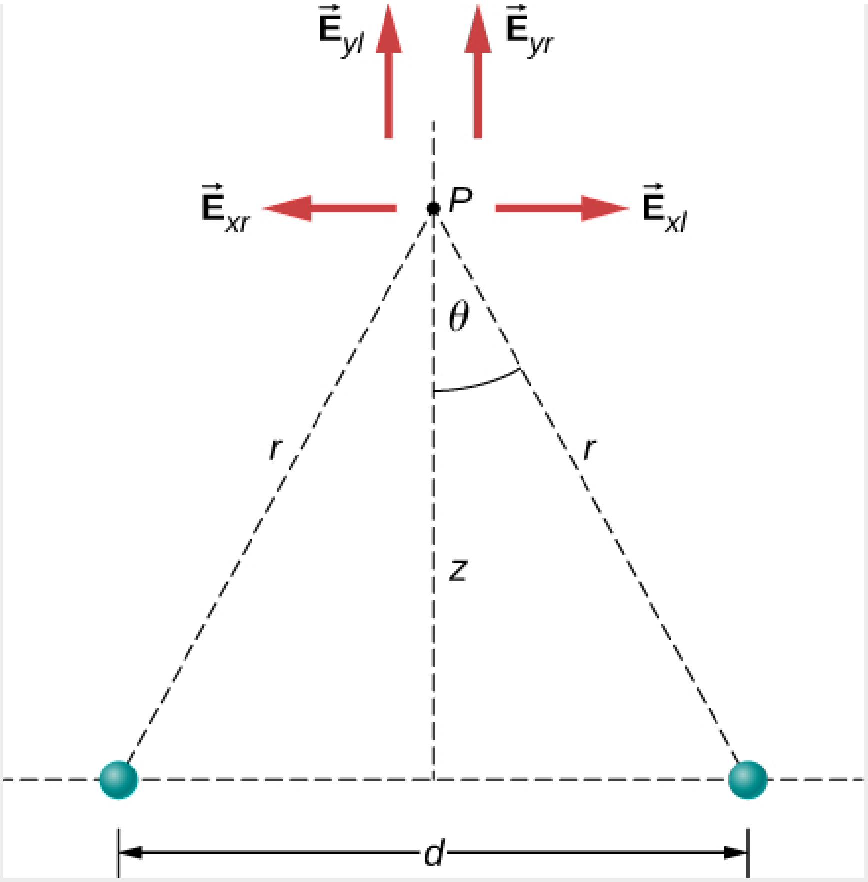


Finding the field of two identical source charges at the point *P*. Due to the symmetry, the net field at *P* is entirely vertical. (Notice that this is *not* true away from the midline between the charges.)

Strategy
We add the two fields as vectors, per [link]. Notice that the system (and therefore the field) is symmetrical about the vertical axis; as a result, the horizontal components of the field vectors cancel. This simplifies the math. Also, we take care to express our final answer in terms of only quantities that are given in the original statement of the problem: *q*, *z*, *d*, and constants (π, ε_0).

Solution

a. By symmetry, the horizontal (*x*)-components of $\vec{\mathbf{E}}$ cancel ([link]);
$E_x = \frac{1}{4\pi\varepsilon_0}\,\frac{q}{r^2}\sin\theta - \frac{1}{4\pi\varepsilon_0}\,\frac{q}{r^2}\sin\theta = 0.$



Note that the horizontal components of the electric fields from the two charges cancel each other out, while the vertical components add together.

The vertical (z)-component is given by

Equation:

$$E_z = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\cos\theta + \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\cos\theta = \frac{1}{4\pi\varepsilon_0}\frac{2q}{r^2}\cos\theta.$$

Since none of the other components survive, this is the entire electric field, and it points in the $\widehat{\mathbf{k}}$ direction. Notice that this calculation uses the principle of **superposition**; we calculate the fields of the two charges independently and then add them together.

What we want to do now is replace the quantities in this expression that we don’t know (such as r), or can’t easily measure (such as $\cos\theta$) with quantities that we do know, or can measure. In this case, by geometry,

Equation:

$$r^2 = z^2 + \left(\frac{d}{2}\right)^2$$

and

Equation:

$$\cos\theta = \frac{z}{r} = \frac{z}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{1/2}}.$$

Thus, substituting,

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0} \frac{2q}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]} \frac{z}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{1/2}}\hat{\mathbf{k}}.$$

Simplifying, the desired answer is

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0} \frac{2qz}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{3/2}}\hat{\mathbf{k}}.$$

b. If the source charges are equal and opposite, the vertical components cancel because $E_z = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\cos\theta - \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\cos\theta = 0$

and we get, for the horizontal component of $\vec{\mathbf{E}}$,

Equation:

$$\begin{aligned}\vec{\mathbf{E}}(z) &= \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\sin\theta\hat{\mathbf{i}} - \frac{1}{4\pi\varepsilon_0}\frac{-q}{r^2}\sin\theta\hat{\mathbf{i}} \\ &= \frac{1}{4\pi\varepsilon_0}\frac{2q}{r^2}\sin\theta\hat{\mathbf{i}} \\ &= \frac{1}{4\pi\varepsilon_0}\frac{2q}{\left[z^2+\left(\frac{d}{2}\right)^2\right]}\frac{\left(\frac{d}{2}\right)}{\left[z^2+\left(\frac{d}{2}\right)^2\right]^{1/2}}\hat{\mathbf{i}}.\end{aligned}$$

This becomes

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\frac{qd}{\left[z^2+\left(\frac{d}{2}\right)^2\right]^{3/2}}\hat{\mathbf{i}}.$$

Significance

It is a very common and very useful technique in physics to check whether your answer is reasonable by evaluating it at extreme cases. In this example, we should evaluate the field expressions for the cases $d = 0$, $z \gg d$, and $z \to \infty$, and confirm that the resulting expressions match our physical expectations. Let’s do so:

Let’s start with [link], the field of two identical charges. From far away (i.e., $z \gg d$), the two source charges should “merge” and we should then “see” the field of just one charge, of size $2q$. So, let $z \gg d$; then we can neglect d^2 in [link] to obtain

Equation:

$$\begin{aligned}\lim_{d\to 0}\vec{\mathbf{E}} &= \frac{1}{4\pi\varepsilon_0}\frac{2qz}{\left[z^2\right]^{3/2}}\widehat{\mathbf{k}} \\ &= \frac{1}{4\pi\varepsilon_0}\frac{2qz}{z^3}\widehat{\mathbf{k}} \\ &= \frac{1}{4\pi\varepsilon_0}\frac{(2q)}{z^2}\widehat{\mathbf{k}},\end{aligned}$$

which is the correct expression for a field at a distance z away from a charge $2q$.

Next, we consider the field of equal and opposite charges, [link]. It can be shown (via a Taylor expansion) that for $d \ll z \ll \infty$, this becomes

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\frac{qd}{z^3}\hat{\mathbf{i}},$$

which is the field of a dipole, a system that we will study in more detail later. (Note that the units of $\vec{\mathbf{E}}$ are still correct in this expression, since the units of d in the numerator cancel the unit of the “extra” z in the denominator.) If z is *very* large $(z \to \infty)$, then $E \to 0$, as it should; the two charges “merge” and so cancel out.

Note:
Exercise:

Problem:

Check Your Understanding What is the electric field due to a single point particle?

Solution:

$$\vec{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{r}$$

Note:
Try this simulation of electric field hockey to get the charge in the goal by placing other charges on the field.

Summary

- The electric field is an alteration of space caused by the presence of an electric charge. The electric field mediates the electric force between a source charge and a test charge.

- The electric field, like the electric force, obeys the superposition principle
- The field is a vector; by definition, it points away from positive charges and toward negative charges.

Conceptual Questions

Exercise:

Problem:

When measuring an electric field, could we use a negative rather than a positive test charge?

Solution:

Either sign of the test charge could be used, but the convention is to use a positive test charge.

Exercise:

Problem:

During fair weather, the electric field due to the net charge on Earth points downward. Is Earth charged positively or negatively?

Exercise:

Problem:

If the electric field at a point on the line between two charges is zero, what do you know about the charges?

Solution:

The charges are of the same sign.

Exercise:

Problem:

Two charges lie along the *x*-axis. Is it true that the net electric field always vanishes at some point (other than infinity) along the *x*-axis?

Problems

Exercise:

Problem:

A particle of charge $2.0\ \times\ 10^{-8}$ C experiences an upward force of magnitude $4.0\ \times\ 10^{-6}$ N when it is placed in a particular point in an electric field. (a) What is the electric field at that point? (b) If a charge $q=-1.0\ \times\ 10^{-8}$ C is placed there, what is the force on it?

Solution:

a. $E=2.0\ \times\ 10^{-2}\,\frac{\text{N}}{\text{C}}$ up;
b. $F=2.0\ \times\ 10^{-6}$ N down

Exercise:

Problem:

On a typical clear day, the atmospheric electric field points downward and has a magnitude of approximately 100 N/C. Compare the gravitational and electric forces on a small dust particle of mass $2.0\ \times\ 10^{-15}$ g that carries a single electron charge. What is the acceleration (both magnitude and direction) of the dust particle?

Exercise:

Problem:

Consider an electron that is $10^{-10}\ \mathrm{m}$ from an alpha particle $(q = 3.2\ \times\ 10^{-19}\ \mathrm{C})$. (a) What is the electric field due to the alpha particle at the location of the electron? (b) What is the electric field due to the electron at the location of the alpha particle? (c) What is the electric force on the alpha particle? On the electron?

Solution:

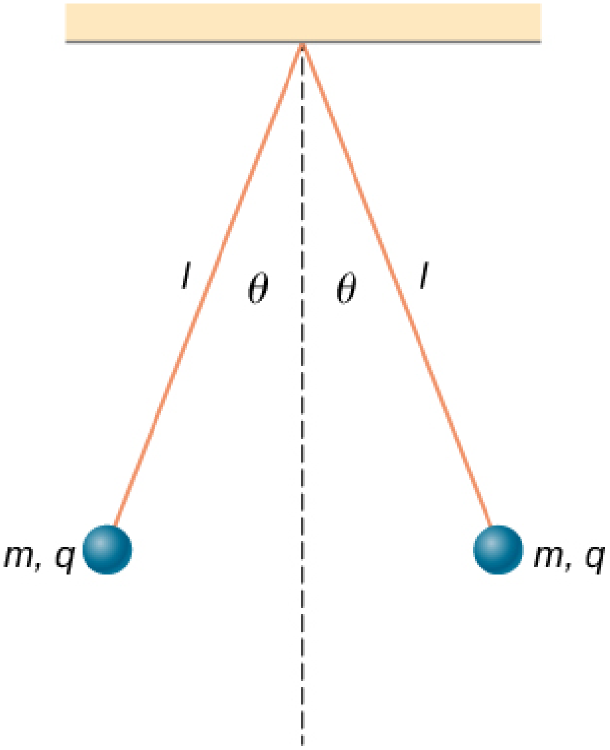
a. $E = 2.88\ \times\ 10^{11}\ \mathrm{N/C}$;
b. $E = 1.44\ \times\ 10^{11}\ \mathrm{N/C}$;
c. $F = 4.61\ \times\ 10^{-8}\ \mathrm{N}$ on alpha particle;
$F = 4.61\ \times\ 10^{-8}\ \mathrm{N}$ on electron

Exercise:

Problem:

Each the balls shown below carries a charge q and has a mass m. The length of each thread is l, and at equilibrium, the balls are separated by an angle 2θ. How does θ vary with q and l? Show that θ satisfies

$$\sin(\theta)^2 \tan(\theta) = \frac{q^2}{16\pi\varepsilon_0 g l^2 m}.$$



Exercise:

Problem:

What is the electric field at a point where the force on a -2.0×10^{-6} –C charge is $\left(4.0\hat{\mathbf{i}} - 6.0\hat{\mathbf{j}}\right) \times 10^{-6}$ N?

Solution:

$$E = \left(-2.0\hat{\mathbf{i}} + 3.0\hat{\mathbf{j}}\right)\text{ N}$$

Exercise:

Problem:

A proton is suspended in the air by an electric field at the surface of Earth. What is the strength of this electric field?

Exercise:

Problem:

The electric field in a particular thundercloud is $2.0\;\times\;10^{5}\;\mathrm{N/C}$. What is the acceleration of an electron in this field?

Solution:

$F=3.204\;\times\;10^{-14}\;\mathrm{N}$,
$a=3.517\;\times\;10^{16}\;\mathrm{m/s}^{2}$

Exercise:

Problem:

A small piece of cork whose mass is 2.0 g is given a charge of $5.0\;\times\;10^{-7}\;\mathrm{C}$. What electric field is needed to place the cork in equilibrium under the combined electric and gravitational forces?

Exercise:

Problem:

If the electric field is $100\;\mathrm{N/C}$ at a distance of 50 cm from a point charge q, what is the value of q?

Solution:

$q=2.78\;\times\;10^{-9}\;\mathrm{C}$

Exercise:

Problem:

What is the electric field of a proton at the first Bohr orbit for hydrogen $(r=5.29\;\times\;10^{-11}\;\mathrm{m})$? What is the force on the electron in that orbit?

Exercise:

Problem:

(a) What is the electric field of an oxygen nucleus at a point that is 10^{-10} m from the nucleus? (b) What is the force this electric field exerts on a second oxygen nucleus placed at that point?

Solution:

a. $E = 1.15 \times 10^{12}\,\text{N/C}$;
b. $F = 1.47 \times 10^{-6}\,\text{N}$

Exercise:

Problem:

Two point charges, $q_1 = 2.0 \times 10^{-7}$ C and $q_2 = -6.0 \times 10^{-8}$ C, are held 25.0 cm apart. (a) What is the electric field at a point 5.0 cm from the negative charge and along the line between the two charges? (b)What is the force on an electron placed at that point?

Exercise:

Problem:

Point charges $q_1 = 50\,\mu\text{C}$ and $q_2 = -25\,\mu\text{C}$ are placed 1.0 m apart. (a) What is the electric field at a point midway between them? (b) What is the force on a charge $q_3 = 20\,\mu\text{C}$ situated there?

Solution:

If the q_2 is to the right of q_1, the electric field vector from both charges point to the right. a. $E = 2.70 \times 10^{6}\,\text{N/C}$;
b. $F = 54.0\,\text{N}$

Exercise:

Problem:

Can you arrange the two point charges $q_1 = -2.0 \times 10^{-6}$ C and $q_2 = 4.0 \times 10^{-6}$ C along the *x*-axis so that $E = 0$ at the origin?

Exercise:

Problem:

Point charges $q_1 = q_2 = 4.0 \times 10^{-6}$ C are fixed on the *x*-axis at $x = -3.0\,\mathrm{m}$ and $x = 3.0\,\mathrm{m}$. What charge *q* must be placed at the origin so that the electric field vanishes at $x = 0, y = 3.0\,\mathrm{m}$?

Solution:

There is 45° right triangle geometry. The *x*-components of the electric field at $y = 3\,\mathrm{m}$ cancel. The *y*-components give
$E(y = 3\,\mathrm{m}) = 2.83 \times 10^3\,\mathrm{N/C}$.
At the origin we have a a negative charge of magnitide
$q = -2.83 \times 10^{-6}$ C.

Glossary

electric field
: physical phenomenon created by a charge; it “transmits” a force between a two charges

superposition
: concept that states that the net electric field of multiple source charges is the vector sum of the field of each source charge calculated individually

Calculating Electric Fields of Charge Distributions

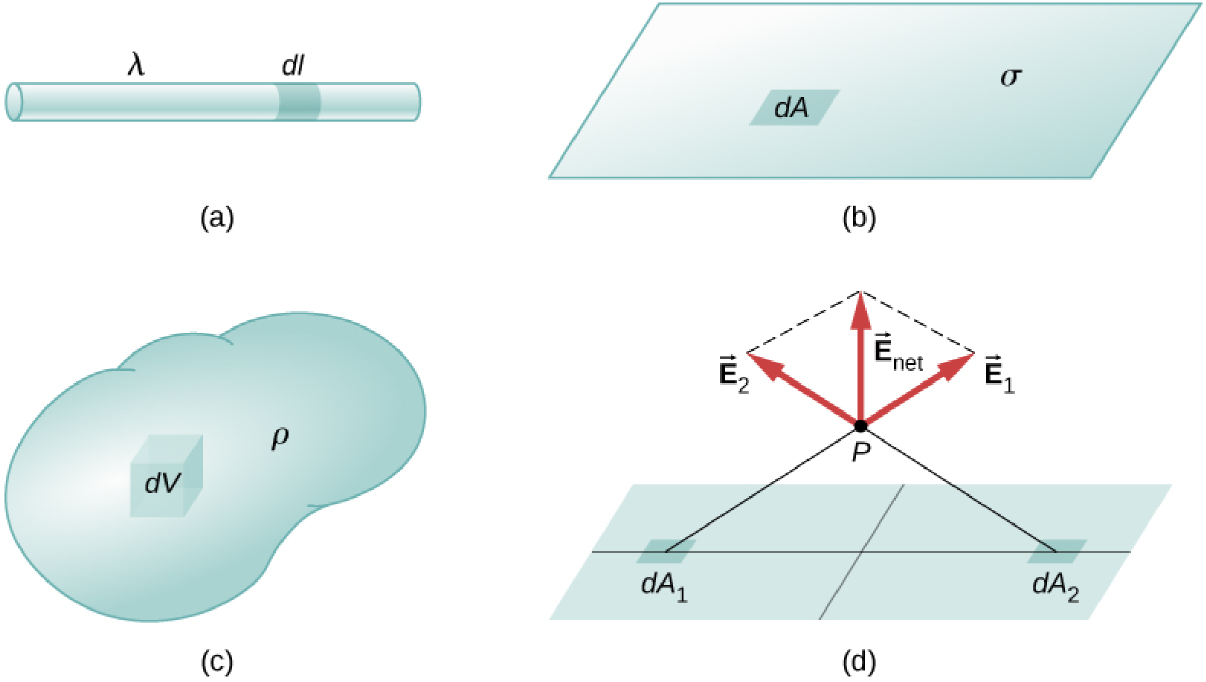
By the end of this section, you will be able to:

- Explain what a continuous source charge distribution is and how it is related to the concept of quantization of charge
- Describe line charges, surface charges, and volume charges
- Calculate the field of a continuous source charge distribution of either sign

The charge distributions we have seen so far have been discrete: made up of individual point particles. This is in contrast with a **continuous charge distribution**, which has at least one nonzero dimension. If a charge distribution is continuous rather than discrete, we can generalize the definition of the electric field. We simply divide the charge into infinitesimal pieces and treat each piece as a point charge.

Note that because charge is quantized, there is no such thing as a “truly” continuous charge distribution. However, in most practical cases, the total charge creating the field involves such a huge number of discrete charges that we can safely ignore the discrete nature of the charge and consider it to be continuous. This is exactly the kind of approximation we make when we deal with a bucket of water as a continuous fluid, rather than a collection of $\mathrm{H_2O}$ molecules.

Our first step is to define a charge density for a charge distribution along a line, across a surface, or within a volume, as shown in [link].



The configuration of charge differential elements for a (a) line charge, (b) sheet of charge, and (c) a volume of charge. Also note that (d) some of the components of the total electric field cancel out, with the remainder resulting in a net electric field.

Definitions of charge density:

- $\lambda \equiv$ charge per unit length (**linear charge density**); units are coulombs per meter (C/m)
- $\sigma \equiv$ charge per unit area (**surface charge density**); units are coulombs per square meter $(\mathrm{C/m^2})$
- $\rho \equiv$ charge per unit volume (**volume charge density**); units are coulombs per cubic meter $(\mathrm{C/m^3})$

Then, for a line charge, a surface charge, and a volume charge, the summation in [link] becomes an integral and q_i is replaced by $dq = \lambda dl$, σdA, or ρdV, respectively:

Equation:

$$\text{Point charges:} \qquad \vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \left(\frac{q_i}{r^2}\right) \hat{\mathbf{r}}$$

Equation:

$$\text{Line charge:} \qquad \vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2}\right) \hat{\mathbf{r}}$$

Equation:

$$\text{Surface charge:} \qquad \vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{surface}} \left(\frac{\sigma dA}{r^2}\right) \hat{\mathbf{r}}$$

Equation:

$$\text{Volume charge:} \qquad \vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{volume}} \left(\frac{\rho dV}{r^2}\right) \hat{\mathbf{r}}$$

The integrals are generalizations of the expression for the field of a point charge. They implicitly include and assume the principle of superposition. The “trick” to using them is almost always in coming up with correct expressions for *dl*, *dA*, or *dV*, as the case may be, expressed in terms of *r*, and also expressing the charge density function appropriately. It may be constant; it might be dependent on location.

Note carefully the meaning of *r* in these equations: It is the distance from the charge element $(q_i, \lambda dl, \sigma dA, \rho dV)$ to the location of interest, $P\left(x, y, z\right)$ (the point in space where you want to determine the field). However, don’t confuse this with the meaning of $\hat{\mathbf{r}}$; we are using it and the vector notation $\vec{\mathbf{E}}$ to write three integrals at once. That is, [link] is actually

Equation:

$$E_x\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2}\right)_x, \quad E_y\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2}\right)_y, \quad E_z\left(P\right) = \frac{1}{4\pi\varepsilon_0} \int_{\text{line}} \left(\frac{\lambda dl}{r^2}\right)_z.$$

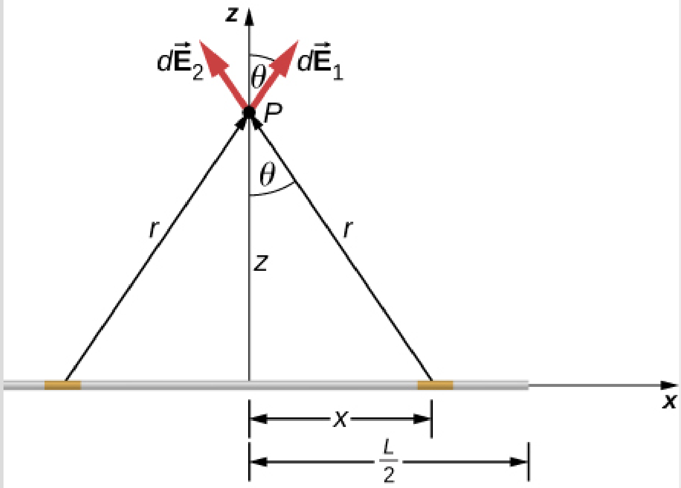
Example:

Electric Field of a Line Segment

Find the electric field a distance z above the midpoint of a straight line segment of length L that carries a uniform line charge density λ.

Strategy

Since this is a continuous charge distribution, we conceptually break the wire segment into differential pieces of length dl, each of which carries a differential amount of charge $dq = \lambda dl$. Then, we calculate the differential field created by two symmetrically placed pieces of the wire, using the symmetry of the setup to simplify the calculation ([link]). Finally, we integrate this differential field expression over the length of the wire (half of it, actually, as we explain below) to obtain the complete electric field expression.



A uniformly charged segment of wire. The electric field at point P can be found by applying the superposition principle to symmetrically placed charge elements and integrating.

Solution

Before we jump into it, what do we expect the field to "look like" from far away? Since it is a finite line segment, from far away, it should look like a point charge. We will check the expression we get to see if it meets this expectation.

The electric field for a line charge is given by the general expression

Equation:

$$\vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0}\int_{\text{line}}\frac{\lambda dl}{r^2}\hat{\mathbf{r}}.$$

The symmetry of the situation (our choice of the two identical differential pieces of charge) implies the horizontal (x)-components of the field cancel, so that the net field points in the z-direction. Let's check this formally.

The total field $\vec{\mathbf{E}}\left(P\right)$ is the vector sum of the fields from each of the two charge elements (call them $\vec{\mathbf{E}}_1$ and $\vec{\mathbf{E}}_2$, for now):

Equation:

$$\vec{\mathbf{E}}(P) = \vec{\mathbf{E}}_1 + \vec{\mathbf{E}}_2 = E_{1x}\hat{\mathbf{i}} + E_{1z}\widehat{\mathbf{k}} + E_{2x}\left(-\hat{\mathbf{i}}\right) + E_{2z}\widehat{\mathbf{k}}.$$

Because the two charge elements are identical and are the same distance away from the point *P* where we want to calculate the field, $E_{1x} = E_{2x}$, so those components cancel. This leaves
Equation:

$$\vec{\mathbf{E}}(P) = E_{1z}\widehat{\mathbf{k}} + E_{2z}\widehat{\mathbf{k}} = E_1 \cos\theta\widehat{\mathbf{k}} + E_2 \cos\theta\widehat{\mathbf{k}}.$$

These components are also equal, so we have
Equation:

$$\begin{aligned} \vec{\mathbf{E}}(P) &= \tfrac{1}{4\pi\varepsilon_0}\int \frac{\lambda dl}{r^2}\cos\theta\widehat{\mathbf{k}} + \tfrac{1}{4\pi\varepsilon_0}\int \frac{\lambda dl}{r^2}\cos\theta\widehat{\mathbf{k}} \\ &= \tfrac{1}{4\pi\varepsilon_0}\int_0^{L/2} \frac{2\lambda dx}{r^2}\cos\theta\widehat{\mathbf{k}} \end{aligned}$$

where our differential line element *dl* is *dx*, in this example, since we are integrating along a line of charge that lies on the *x*-axis. (The limits of integration are 0 to $\frac{L}{2}$, not $-\frac{L}{2}$ to $+\frac{L}{2}$, because we have constructed the net field from two differential pieces of charge *dq*. If we integrated along the entire length, we would pick up an erroneous factor of 2.)

In principle, this is complete. However, to actually calculate this integral, we need to eliminate all the variables that are not given. In this case, both *r* and *θ* change as we integrate outward to the end of the line charge, so those are the variables to get rid of. We can do that the same way we did for the two point charges: by noticing that
Equation:

$$r = \left(z^2 + x^2\right)^{1/2}$$

and
Equation:

$$\cos\theta = \frac{z}{r} = \frac{z}{\left(z^2 + x^2\right)^{1/2}}.$$

Substituting, we obtain
Equation:

$$\begin{aligned} \vec{\mathbf{E}}(P) &= \tfrac{1}{4\pi\varepsilon_0}\int_0^{L/2} \frac{2\lambda dx}{\left(z^2 + x^2\right)} \tfrac{z}{\left(z^2+x^2\right)^{1/2}}\widehat{\mathbf{k}} \\ &= \tfrac{1}{4\pi\varepsilon_0}\int_0^{L/2} \frac{2\lambda z}{\left(z^2 + x^2\right)^{3/2}} dx\widehat{\mathbf{k}} \\ &= \tfrac{2\lambda z}{4\pi\varepsilon_0}\left[\tfrac{x}{z^2\sqrt{z^2+x^2}}\right]\Bigg|_0^{L/2}\widehat{\mathbf{k}} \end{aligned}$$

which simplifies to
Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\,\frac{\lambda L}{z\sqrt{z^2 + \frac{L^2}{4}}}\widehat{\mathbf{k}}.$$

Significance
Notice, once again, the use of symmetry to simplify the problem. This is a very common strategy for calculating electric fields. The fields of nonsymmetrical charge distributions have to be handled with multiple integrals and may need to be calculated numerically by a computer.

Note:
Exercise:

Problem:

Check Your Understanding How would the strategy used above change to calculate the electric field at a point a distance *z* above one end of the finite line segment?

Solution:

We will no longer be able to take advantage of symmetry. Instead, we will need to calculate each of the two components of the electric field with their own integral.

Example:
Electric Field of an Infinite Line of Charge
Find the electric field a distance *z* above the midpoint of an infinite line of charge that carries a uniform line charge density λ.
Strategy
This is exactly like the preceding example, except the limits of integration will be $-\infty$ to $+\infty$.
Solution
Again, the horizontal components cancel out, so we wind up with
Equation:

$$\vec{\mathbf{E}}\left(P\right) = \frac{1}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\lambda dx}{r^2}\cos\theta\widehat{\mathbf{k}}$$

where our differential line element *dl* is *dx*, in this example, since we are integrating along a line of charge that lies on the *x*-axis. Again,
Equation:

$$\cos\theta = \frac{z}{r} = \frac{z}{\left(z^2 + x^2\right)^{1/2}}.$$

Substituting, we obtain
Equation:

$$\begin{aligned}\vec{\mathbf{E}}(P) &= \frac{1}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\lambda dx}{(z^2+x^2)}\frac{z}{(z^2+x^2)^{1/2}}\widehat{\mathbf{k}}\\ &= \frac{1}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\lambda z}{(z^2+x^2)^{3/2}}dx\widehat{\mathbf{k}}\\ &= \frac{\lambda z}{4\pi\varepsilon_0}\left[\frac{x}{z^2\sqrt{z^2+x^2}}\right]\Big|_{-\infty}^{\infty}\widehat{\mathbf{k}},\end{aligned}$$

which simplifies to

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\widehat{\mathbf{k}}.$$

Significance

Our strategy for working with continuous charge distributions also gives useful results for charges with infinite dimension.

In the case of a finite line of charge, note that for $z \gg L$, z^2 dominates the L in the denominator, so that [link] simplifies to

Equation:

$$\vec{\mathbf{E}} \approx \frac{1}{4\pi\varepsilon_0}\frac{\lambda L}{z^2}\widehat{\mathbf{k}}.$$

If you recall that $\lambda L = q$, the total charge on the wire, we have retrieved the expression for the field of a point charge, as expected.

In the limit $L \to \infty$, on the other hand, we get the field of an **infinite straight wire**, which is a straight wire whose length is much, much greater than either of its other dimensions, and also much, much greater than the distance at which the field is to be calculated:

Note:

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\widehat{\mathbf{k}}.$$

An interesting artifact of this infinite limit is that we have lost the usual $1/r^2$ dependence that we are used to. This will become even more intriguing in the case of an infinite plane.

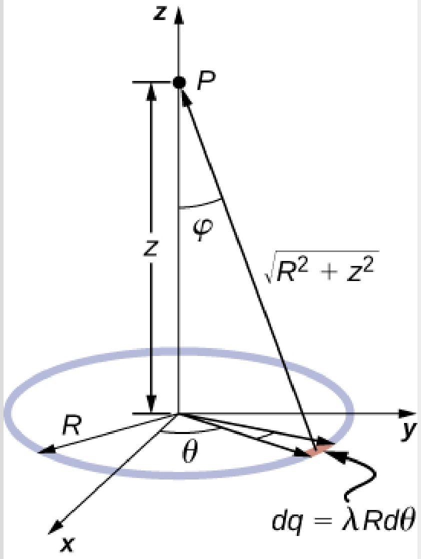
Example:

Electric Field due to a Ring of Charge

A ring has a uniform charge density λ, with units of coulomb per unit meter of arc. Find the electric field at a point on the axis passing through the center of the ring.

Strategy

We use the same procedure as for the charged wire. The difference here is that the charge is distributed on a circle. We divide the circle into infinitesimal elements shaped as arcs on the circle and use polar coordinates shown in [link].



The system and variable for calculating the electric field due to a ring of charge.

Solution

The electric field for a line charge is given by the general expression

Equation:

$$\vec{\mathbf{E}}(P) = \frac{1}{4\pi\varepsilon_0} \int_{\text{line}} \frac{\lambda dl}{r^2} \hat{\mathbf{r}}.$$

A general element of the arc between θ and $\theta + d\theta$ is of length $Rd\theta$ and therefore contains a charge equal to $\lambda Rd\theta$. The element is at a distance of $r = \sqrt{z^2 + R^2}$ from P, the angle is $\cos\phi = \frac{z}{\sqrt{z^2+R^2}}$, and therefore the electric field is

Equation:

$$\begin{aligned}\vec{\mathbf{E}}(P) &= \frac{1}{4\pi\varepsilon_0}\int_{\textbf{line}}\frac{\lambda dl}{r^2}\hat{\mathbf{r}} = \frac{1}{4\pi\varepsilon_0}\int_0^{2\pi}\frac{\lambda R d\theta}{z^2+R^2}\frac{z}{\sqrt{z^2+R^2}}\widehat{\mathbf{z}}\\ &= \frac{1}{4\pi\varepsilon_0}\frac{\lambda R z}{\left(z^2+R^2\right)^{3/2}}\widehat{\mathbf{z}}\int_0^{2\pi} d\theta = \frac{1}{4\pi\varepsilon_0}\frac{2\pi\lambda R z}{\left(z^2+R^2\right)^{3/2}}\widehat{\mathbf{z}}\\ &= \frac{1}{4\pi\varepsilon_0}\frac{q_{\text{tot}} z}{\left(z^2+R^2\right)^{3/2}}\widehat{\mathbf{z}}.\end{aligned}$$

Significance

As usual, symmetry simplified this problem, in this particular case resulting in a trivial integral. Also, when we take the limit of $z >> R$, we find that

Equation:

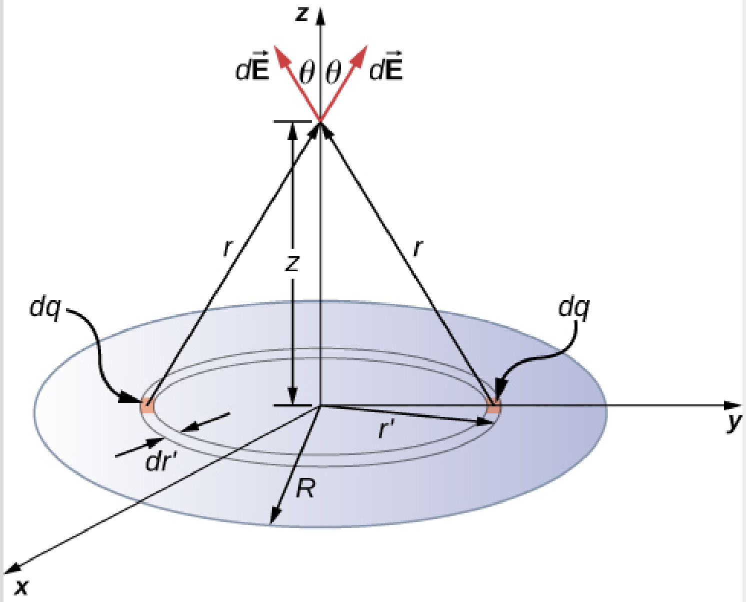
$$\vec{\mathbf{E}} \approx \frac{1}{4\pi\varepsilon_0}\frac{q_{\text{tot}}}{z^2}\widehat{\mathbf{z}},$$

as we expect.

Example:

The Field of a Disk

Find the electric field of a circular thin disk of radius *R* and uniform charge density at a distance *z* above the center of the disk ([link])



A uniformly charged disk. As in the line charge example, the field above the center of this disk can be calculated by taking advantage of the symmetry of the charge distribution.

Strategy

The electric field for a surface charge is given by

Equation:

$$\vec{\mathbf{E}}(P)=\frac{1}{4\pi\varepsilon_0}\int_{\text{surface}}\frac{\sigma dA}{r^2}\hat{\mathbf{r}}.$$

To solve surface charge problems, we break the surface into symmetrical differential “stripes” that match the shape of the surface; here, we’ll use rings, as shown in the figure. Again, by symmetry, the horizontal components cancel and the field is entirely in the vertical $(\widehat{\mathbf{k}})$ direction. The vertical component of the electric field is extracted by multiplying by $\cos\theta$, so

Equation:

$$\vec{\mathbf{E}}(P)=\frac{1}{4\pi\varepsilon_0}\int_{\text{surface}}\frac{\sigma dA}{r^2}\cos\theta\,\widehat{\mathbf{k}}.$$

As before, we need to rewrite the unknown factors in the integrand in terms of the given quantities. In this case,

Equation:

$$\begin{aligned} dA &= 2\pi r' dr' \\ r^2 &= {r'}^2+z^2 \\ \cos\theta &= \tfrac{z}{\left({r'}^2+z^2\right)^{1/2}}. \end{aligned}$$

(Please take note of the two different “r’s” here; r is the distance from the differential ring of charge to the point P where we wish to determine the field, whereas r' is the distance from the center of the disk to the differential ring of charge.) Also, we already performed the polar angle integral in writing down dA.

Solution

Substituting all this in, we get

Equation:

$$\begin{aligned} \vec{\mathbf{E}}(P) &= \vec{\mathbf{E}}\left(z\right)=\tfrac{1}{4\pi\varepsilon_0}\int_0^R\frac{\sigma\left(2\pi r' dr'\right)z}{\left({r'}^2+z^2\right)^{3/2}}\widehat{\mathbf{k}} \\ &= \tfrac{1}{4\pi\varepsilon_0}(2\pi\sigma z)\left(\tfrac{1}{z}-\tfrac{1}{\sqrt{R^2+z^2}}\right)\widehat{\mathbf{k}} \end{aligned}$$

or, more simply,

Equation:

$$\vec{\mathbf{E}}(z)=\frac{1}{4\pi\varepsilon_0}\left(2\pi\sigma-\frac{2\pi\sigma z}{\sqrt{R^2+z^2}}\right)\widehat{\mathbf{k}}.$$

Significance

Again, it can be shown (via a Taylor expansion) that when $z\gg R$, this reduces to

Equation:

$$\vec{\mathbf{E}}(z) \approx \frac{1}{4\pi\varepsilon_0}\frac{\sigma\pi R^2}{z^2}\widehat{\mathbf{k}},$$

which is the expression for a point charge $Q = \sigma\pi R^2$.

Note:
Exercise:

Problem:

Check Your Understanding How would the above limit change with a uniformly charged rectangle instead of a disk?

Solution:

The point charge would be $Q = \sigma ab$ where a and b are the sides of the rectangle but otherwise identical.

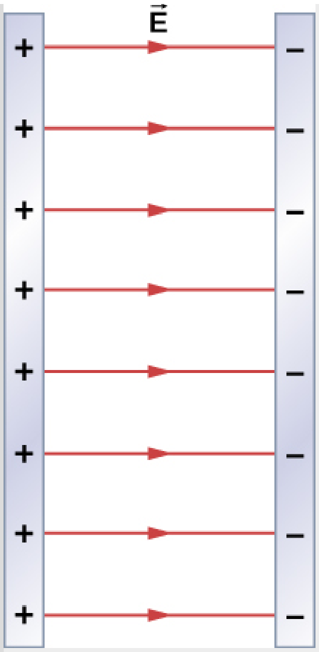
As $R \rightarrow \infty$, [link] reduces to the field of an **infinite plane**, which is a flat sheet whose area is much, much greater than its thickness, and also much, much greater than the distance at which the field is to be calculated:

Note:
Equation:

$$\vec{\mathbf{E}} = \frac{\sigma}{2\varepsilon_0}\widehat{\mathbf{k}}.$$

Note that this field is constant. This surprising result is, again, an artifact of our limit, although one that we will make use of repeatedly in the future. To understand why this happens, imagine being placed above an infinite plane of constant charge. Does the plane look any different if you vary your altitude? No—you still see the plane going off to infinity, no matter how far you are from it. It is important to note that [link] is because we are above the plane. If we were below, the field would point in the $-\widehat{\mathbf{k}}$ direction.

Example:
The Field of Two Infinite Planes
Find the electric field everywhere resulting from two infinite planes with equal but opposite charge densities ([link]).



Two charged infinite planes. Note the direction of the electric field.

Strategy

We already know the electric field resulting from a single infinite plane, so we may use the principle of superposition to find the field from two.

Solution

The electric field points away from the positively charged plane and toward the negatively charged plane. Since the σ are equal and opposite, this means that in the region outside of the two planes, the electric fields cancel each other out to zero.

However, in the region between the planes, the electric fields add, and we get

Equation:

$$\vec{\mathbf{E}} = \frac{\sigma}{\varepsilon_0}\hat{\mathbf{i}}$$

for the electric field. The $\hat{\mathbf{i}}$ is because in the figure, the field is pointing in the +*x*-direction.

Significance

Systems that may be approximated as two infinite planes of this sort provide a useful means of creating uniform electric fields.

Note:

Exercise:

Problem:

Check Your Understanding What would the electric field look like in a system with two parallel positively charged planes with equal charge densities?

Solution:

The electric field would be zero in between, and have magnitude $\frac{\sigma}{\varepsilon_0}$ everywhere else.

Summary

- A very large number of charges can be treated as a continuous charge distribution, where the calculation of the field requires integration. Common cases are:
    - one-dimensional (like a wire); uses a line charge density λ
    - two-dimensional (metal plate); uses surface charge density σ
    - three-dimensional (metal sphere); uses volume charge density ρ
- The "source charge" is a differential amount of charge dq. Calculating dq depends on the type of source charge distribution:
  Equation:

$$dq = \lambda dl; \;\; dq = \sigma dA; \;\; dq = \rho dV.$$

- Symmetry of the charge distribution is usually key.
- Important special cases are the field of an "infinite" wire and the field of an "infinite" plane.

Conceptual Questions

Exercise:

Problem:

Give a plausible argument as to why the electric field outside an infinite charged sheet is constant.

Solution:

At infinity, we would expect the field to go to zero, but because the sheet is infinite in extent, this is not the case. Everywhere you are, you see an infinite plane in all directions.

Exercise:

Problem:

Compare the electric fields of an infinite sheet of charge, an infinite, charged conducting plate, and infinite, oppositely charged parallel plates.

Exercise:

Problem:

Describe the electric fields of an infinite charged plate and of two infinite, charged parallel plates in terms of the electric field of an infinite sheet of charge.

Solution:

The infinite charged plate would have $E = \frac{\sigma}{2\varepsilon_0}$ everywhere. The field would point toward the plate if it were negatively charged and point away from the plate if it were positively charged. The electric field of the parallel plates would be zero between them if they had the same charge, and E would be $E = \frac{\sigma}{\varepsilon_0}$ everywhere else. If the charges were opposite, the situation is reversed, zero outside the plates and $E = \frac{\sigma}{\varepsilon_0}$ between them.

Exercise:

Problem:

A negative charge is placed at the center of a ring of uniform positive charge. What is the motion (if any) of the charge? What if the charge were placed at a point on the axis of the ring other than the center?

Problems

Exercise:

Problem:

A thin conducting plate 1.0 m on the side is given a charge of -2.0×10^{-6} C. An electron is placed 1.0 cm above the center of the plate. What is the acceleration of the electron?

Exercise:

Problem:

Calculate the magnitude and direction of the electric field 2.0 m from a long wire that is charged uniformly at $\lambda = 4.0 \times 10^{-6}\ \mathrm{C/m}$.

Solution:

$\vec{\mathbf{E}}(z) = 3.6 \times 10^4\ \mathrm{N/C}\widehat{\mathbf{k}}$

Exercise:

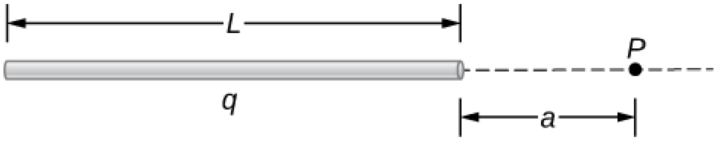
Problem:

Two thin conducting plates, each 25.0 cm on a side, are situated parallel to one another and 5.0 mm apart. If 10^{11} electrons are moved from one plate to the other, what is the electric field between the plates?

Exercise:

Problem:

The charge per unit length on the thin rod shown below is λ. What is the electric field at the point *P*? (*Hint*: Solve this problem by first considering the electric field $d\vec{\mathbf{E}}$ at *P* due to a small segment *dx* of the rod, which contains charge $dq = \lambda dx$. Then find the net field by integrating $d\vec{\mathbf{E}}$ over the length of the rod.)



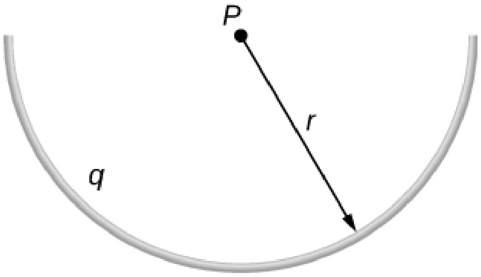
Solution:

$dE = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x+a)^2},\quad E = \frac{\lambda}{4\pi\varepsilon_0}\left[\frac{1}{l+a} - \frac{1}{a}\right]$

Exercise:

Problem:

The charge per unit length on the thin semicircular wire shown below is λ. What is the electric field at the point *P*?



Exercise:

Problem:

Two thin parallel conducting plates are placed 2.0 cm apart. Each plate is 2.0 cm on a side; one plate carries a net charge of $8.0\,\mu\text{C}$, and the other plate carries a net charge of $-8.0\,\mu\text{C}$. What is the charge density on the inside surface of each plate? What is the electric field between the plates?

Solution:

$\sigma = 0.02\,\text{C/m}^2 \;\; E = 2.26\,\times\,10^9\,\text{N/C}$

Exercise:

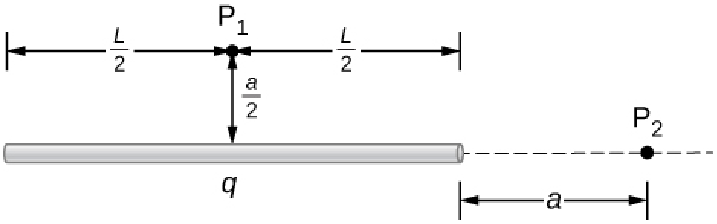
Problem:

A thin conducting plate 2.0 m on a side is given a total charge of $-10.0\ \mu\text{C}$. (a) What is the electric field 1.0 cm above the plate? (b) What is the force on an electron at this point? (c) Repeat these calculations for a point 2.0 cm above the plate. (d) When the electron moves from 1.0 to 2,0 cm above the plate, how much work is done on it by the electric field?

Exercise:

Problem:

A total charge *q* is distributed uniformly along a thin, straight rod of length L (see below). What is the electric field at P_1? At P_2?



Solution:

At P_1: $\vec{\mathbf{E}}(y) = \frac{1}{4\pi\varepsilon_0}\frac{\lambda L}{y\sqrt{y^2+\frac{L^2}{4}}}\hat{\mathbf{j}} \Rightarrow \frac{1}{4\pi\varepsilon_0}\frac{q}{\frac{a}{2}\sqrt{\left(\frac{a}{2}\right)^2+\frac{L^2}{4}}}\hat{\mathbf{j}} = \frac{1}{\pi\varepsilon_0}\frac{q}{a\sqrt{a^2+L^2}}\hat{\mathbf{j}}$

At P_2: Put the origin at the end of L.

$dE = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x+a)^2}, \quad \vec{\mathbf{E}} = -\frac{q}{4\pi\varepsilon_0 l}\left[\frac{1}{l+a} - \frac{1}{a}\right]\hat{\mathbf{i}}$

Exercise:

Problem:

Charge is distributed along the entire *x*-axis with uniform density λ. How much work does the electric field of this charge distribution do on an electron that moves along the *y*-axis from $y = a$ to $y = b$?

Exercise:

Problem:

Charge is distributed along the entire *x*-axis with uniform density λ_x and along the entire *y*-axis with uniform density λ_y. Calculate the resulting electric field at (a) $\vec{\mathbf{r}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ and (b) $\vec{\mathbf{r}} = c\widehat{\mathbf{k}}$.

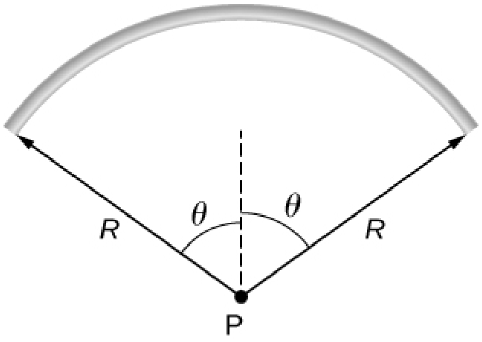
Solution:

a. $\vec{\mathbf{E}}(\vec{\mathbf{r}}) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda_x}{b}\hat{\mathbf{i}} + \frac{1}{4\pi\varepsilon_0}\frac{2\lambda_y}{a}\hat{\mathbf{j}}$; b. $\frac{1}{4\pi\varepsilon_0}\frac{2(\lambda_x+\lambda_y)}{c}\widehat{\mathbf{k}}$

Exercise:

Problem:

A rod bent into the arc of a circle subtends an angle 2θ at the center P of the circle (see below). If the rod is charged uniformly with a total charge Q, what is the electric field at P?



Exercise:

Problem:

A proton moves in the electric field $\vec{\mathbf{E}} = 200\hat{\mathbf{i}}\,\mathrm{N/C}$. (a) What are the force on and the acceleration of the proton? (b) Do the same calculation for an electron moving in this field.

Solution:

a. $\vec{\mathbf{F}} = 3.2\ \times\ 10^{-17}\,\mathrm{N}\hat{\mathbf{i}}$,
$\vec{\mathbf{a}} = 1.92\ \times\ 10^{10}\,\mathrm{m/s^2}\hat{\mathbf{i}}$;
b. $\vec{\mathbf{F}} = -3.2\ \times\ 10^{-17}\,\mathrm{N}\hat{\mathbf{i}}$,
$\vec{\mathbf{a}} = -3.51\ \times\ 10^{13}\,\mathrm{m/s^2}\hat{\mathbf{i}}$

Exercise:

Problem:

An electron and a proton, each starting from rest, are accelerated by the same uniform electric field of 200 N/C. Determine the distance and time for each particle to acquire a kinetic energy of $3.2\ \times\ 10^{-16}\ \mathrm{J}$.

Exercise:

Problem:

A spherical water droplet of radius $25\ \mu\mathrm{m}$ carries an excess 250 electrons. What vertical electric field is needed to balance the gravitational force on the droplet at the surface of the earth?

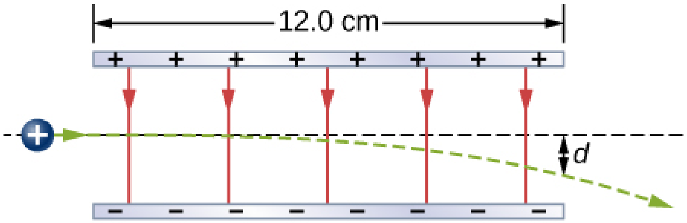
Solution:

$m = 6.5\ \times\ 10^{-11}\,\mathrm{kg}$,
$E = 1.6\ \times\ 10^{7}\,\mathrm{N/C}$

Exercise:

Problem:

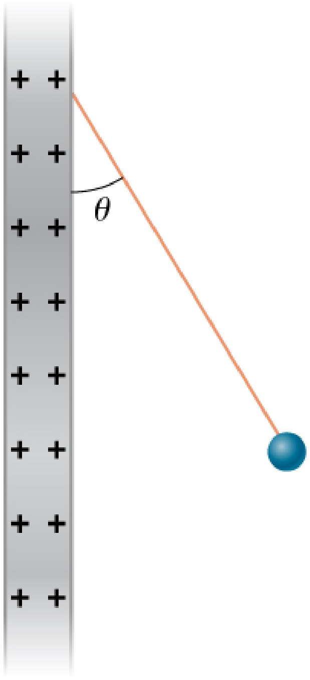
A proton enters the uniform electric field produced by the two charged plates shown below. The magnitude of the electric field is $4.0 \times 10^5\ \mathrm{N/C}$, and the speed of the proton when it enters is $1.5 \times 10^7\ \mathrm{m/s}$. What distance *d* has the proton been deflected downward when it leaves the plates?



Exercise:

Problem:

Shown below is a small sphere of mass 0.25 g that carries a charge of $9.0 \times 10^{-10}\ \mathrm{C}$. The sphere is attached to one end of a very thin silk string 5.0 cm long. The other end of the string is attached to a large vertical conducting plate that has a charge density of $30 \times 10^{-6}\ \mathrm{C/m^2}$. What is the angle that the string makes with the vertical?



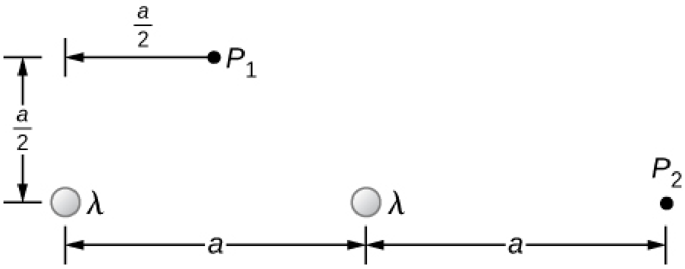
Solution:

$E = 1.70 \times 10^6 \text{ N/C},$
$F = 1.53 \times 10^{-3} \text{ N} \;\; T\cos\theta = mg \;\; T\sin\theta = qE,$
$\tan\theta = 0.62 \Rightarrow \theta = 32.0^\circ,$
This is independent of the length of the string.

Exercise:

Problem:

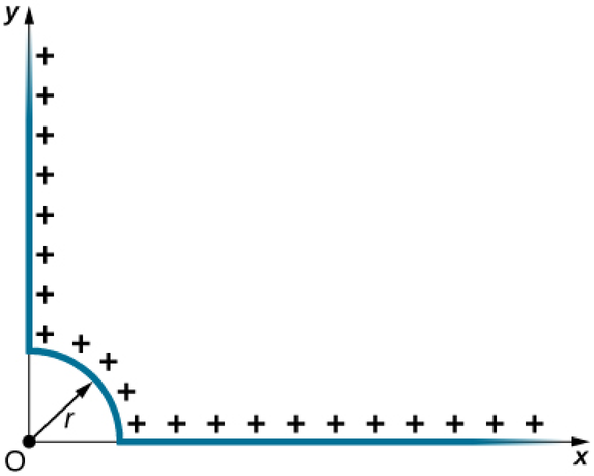
Two infinite rods, each carrying a uniform charge density λ, are parallel to one another and perpendicular to the plane of the page. (See below.) What is the electrical field at P_1? At P_2?



Exercise:

Problem:

Positive charge is distributed with a uniform density λ along the positive *x*-axis from r to ∞, along the positive *y*-axis from r to ∞, and along a 90° arc of a circle of radius *r*, as shown below. What is the electric field at *O*?



Solution:

circular arc $dE_x(-\hat{\mathbf{i}}) = \frac{1}{4\pi\varepsilon_0} \frac{\lambda ds}{r^2} \cos\theta(-\hat{\mathbf{i}}),$
$\vec{\mathbf{E}}_x = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{\mathbf{i}}),$
$dE_y(-\hat{\mathbf{i}}) = \frac{1}{4\pi\varepsilon_0} \frac{\lambda ds}{r^2} \sin\theta(-\hat{\mathbf{j}}),$

$\vec{\mathbf{E}}_y = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{\mathbf{j}});$

y-axis: $\vec{\mathbf{E}}_x = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{\mathbf{i}});$

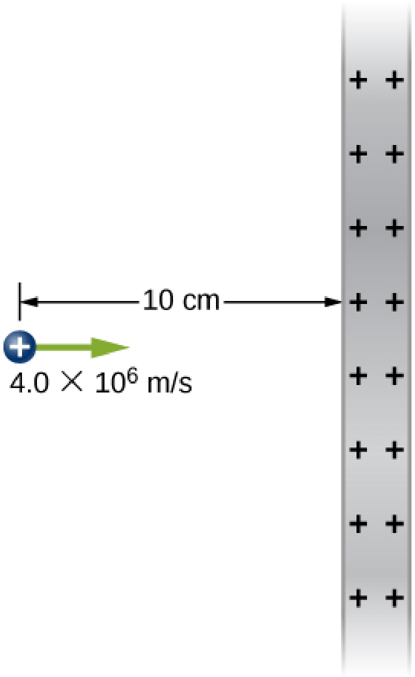
x-axis: $\vec{\mathbf{E}}_y = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{\mathbf{j}}),$

$\vec{\mathbf{E}} = \frac{\lambda}{2\pi\varepsilon_0 r}(-\hat{\mathbf{i}}) + \frac{\lambda}{2\pi\varepsilon_0 r}(-\hat{\mathbf{j}})$

Exercise:

Problem:

From a distance of 10 cm, a proton is projected with a speed of $v = 4.0 \times 10^6\ \mathrm{m/s}$ directly at a large, positively charged plate whose charge density is $\sigma = 2.0 \times 10^{-5}\ \mathrm{C/m^2}$. (See below.) (a) Does the proton reach the plate? (b) If not, how far from the plate does it turn around?



Exercise:

Problem:

A particle of mass m and charge $-q$ moves along a straight line away from a fixed particle of charge Q. When the distance between the two particles is r_0, $-q$ is moving with a speed v_0. (a) Use the work-energy theorem to calculate the maximum separation of the charges. (b) What do you have to assume about v_0 to make this calculation? (c) What is the minimum value of v_0 such that $-q$ escapes from Q?

Solution:

a. $W = \frac{1}{2}m(v^2 - v_0^2)$, $\frac{Qq}{4\pi\varepsilon_0}\left(\frac{1}{r} - \frac{1}{r_0}\right) = \frac{1}{2}m(v^2 - v_0^2) \Rightarrow r_0 - r = \frac{4\pi\varepsilon_0}{Qq}\ \frac{1}{2}r r_0 m(v^2 - v_0^2)$; b. $r_0 - r$ is negative; therefore, $v_0 > v$,

$r \to \infty,$ and $v \to 0$: $\frac{Qq}{4\pi\varepsilon_0}\left(-\frac{1}{r_0}\right) = -\frac{1}{2}mv_0^2 \Rightarrow v_0 = \sqrt{\frac{Qq}{2\pi\varepsilon_0 m r_0}}$

Glossary

continuous charge distribution

total source charge composed of so large a number of elementary charges that it must be treated as continuous, rather than discrete

infinite plane

flat sheet in which the dimensions making up the area are much, much greater than its thickness, and also much, much greater than the distance at which the field is to be calculated; its field is constant

infinite straight wire

straight wire whose length is much, much greater than either of its other dimensions, and also much, much greater than the distance at which the field is to be calculated

linear charge density

amount of charge in an element of a charge distribution that is essentially one-dimensional (the width and height are much, much smaller than its length); its units are C/m

surface charge density

amount of charge in an element of a two-dimensional charge distribution (the thickness is small); its units are $\mathrm{C/m^2}$

volume charge density

amount of charge in an element of a three-dimensional charge distribution; its units are $\mathrm{C/m^3}$

Electric Field Lines

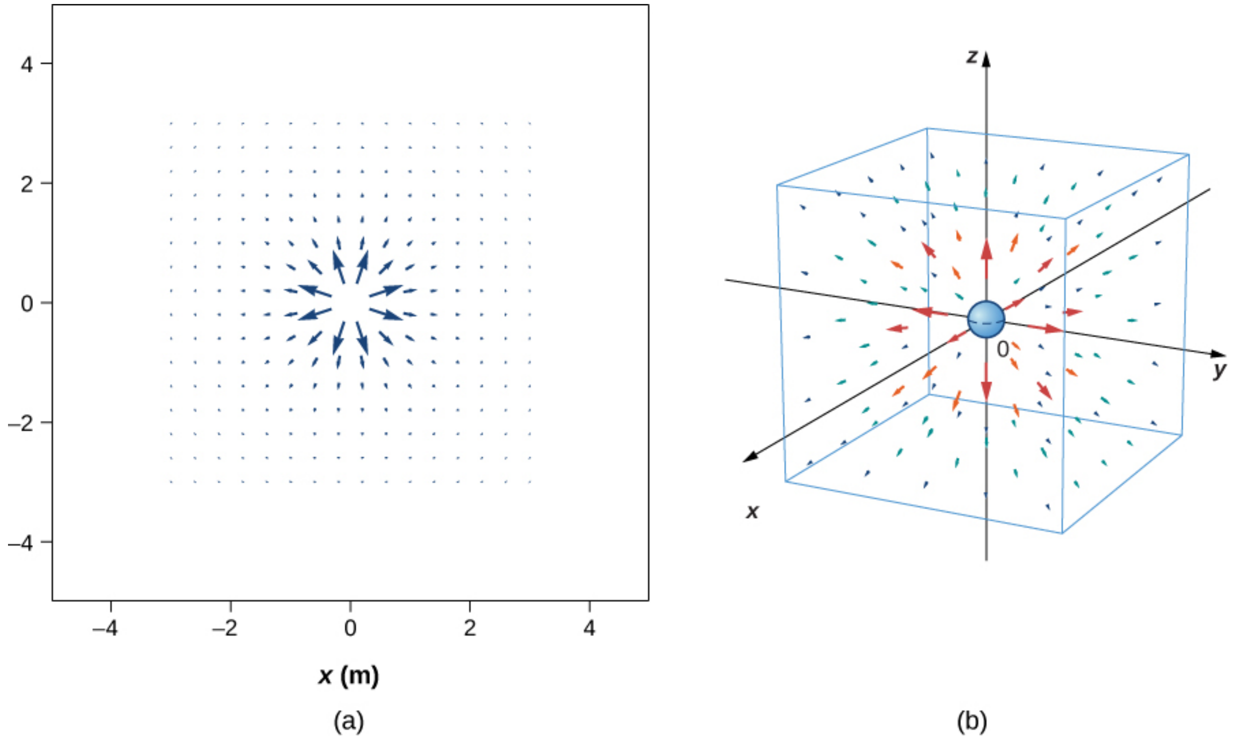
By the end of this section, you will be able to:

- Explain the purpose of an electric field diagram
- Describe the relationship between a vector diagram and a field line diagram
- Explain the rules for creating a field diagram and why these rules make physical sense
- Sketch the field of an arbitrary source charge

Now that we have some experience calculating electric fields, let’s try to gain some insight into the geometry of electric fields. As mentioned earlier, our model is that the charge on an object (the source charge) alters space in the region around it in such a way that when another charged object (the test charge) is placed in that region of space, that test charge experiences an electric force. The concept of electric **field lines**, and of electric field line diagrams, enables us to visualize the way in which the space is altered, allowing us to visualize the field. The purpose of this section is to enable you to create sketches of this geometry, so we will list the specific steps and rules involved in creating an accurate and useful sketch of an electric field.

It is important to remember that electric fields are three-dimensional. Although in this book we include some pseudo-three-dimensional images, several of the diagrams that you’ll see (both here, and in subsequent chapters) will be two-dimensional projections, or cross-sections. Always keep in mind that in fact, you’re looking at a three-dimensional phenomenon.

Our starting point is the physical fact that the electric field of the source charge causes a test charge in that field to experience a force. By definition, electric field vectors point in the same direction as the electric force that a (hypothetical) positive test charge would experience, if placed in the field ([link])



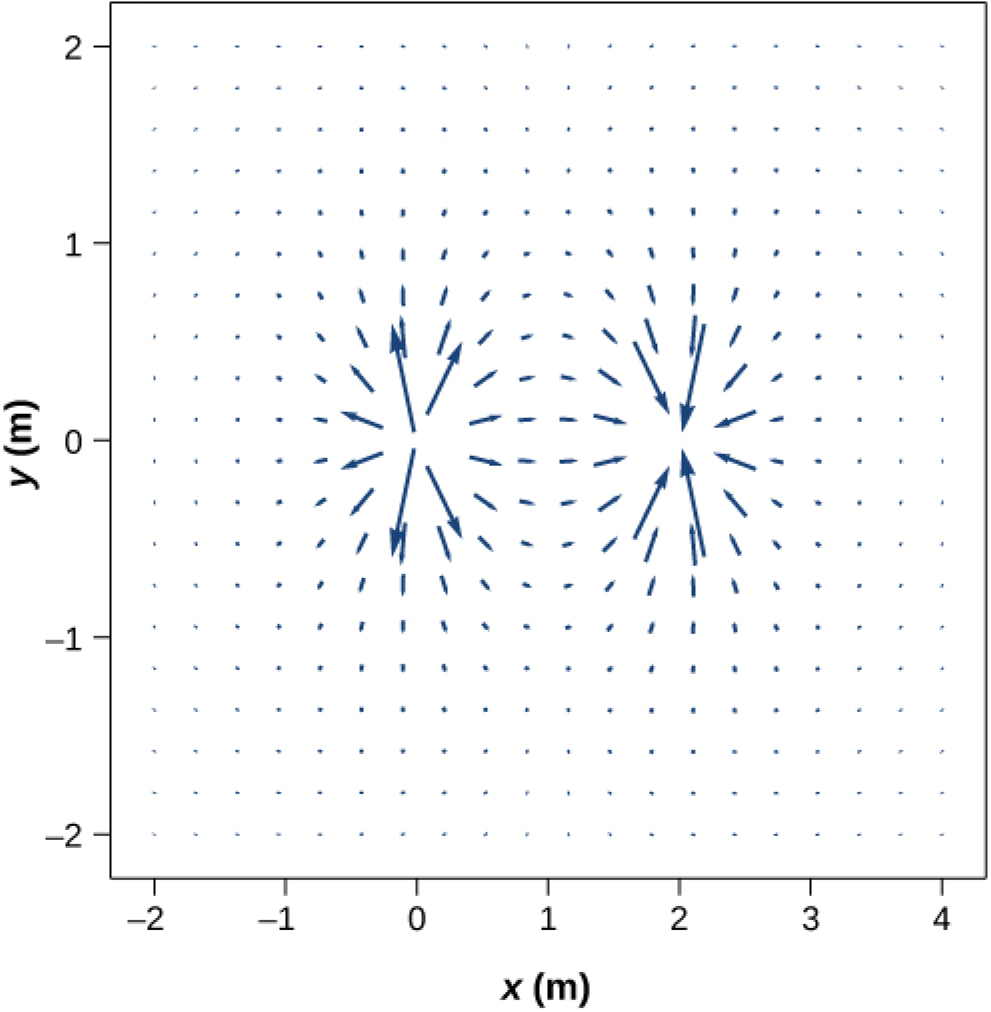
(a)

(b)

The electric field of a positive point charge. A large number of field vectors are shown. Like all vector arrows, the length of each vector is proportional to the magnitude of the field at each point. (a) Field in two dimensions; (b) field in three dimensions.

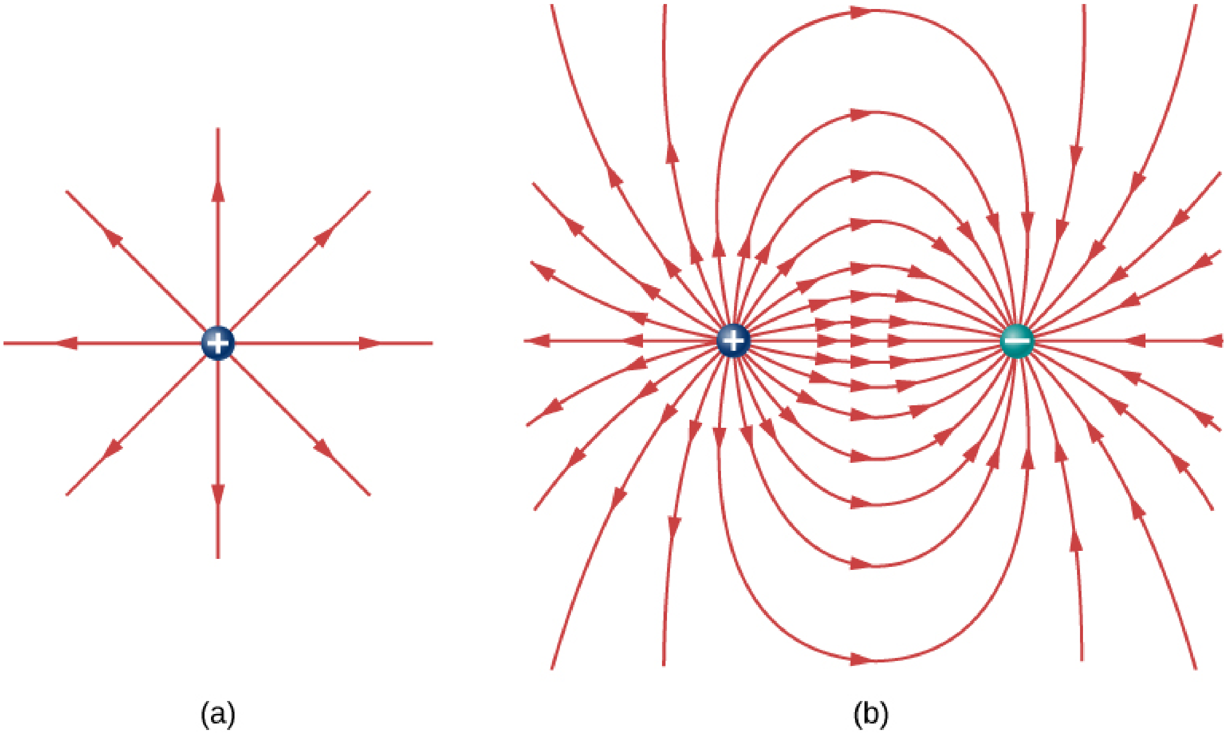
We’ve plotted many field vectors in the figure, which are distributed uniformly around the source charge. Since the electric field is a vector, the arrows that we draw correspond at every point in space to both the magnitude and the direction of the field at that point. As always, the length of the arrow that we draw corresponds to the magnitude of the field vector at that point. For a point source charge, the length decreases by the square of the distance from the source charge. In addition, the direction of the field vector is radially away from the source charge, because the direction of the electric field is defined by the direction of the force that a positive test charge would experience in that field. (Again, keep in mind that the actual field is three-dimensional; there are also field lines pointing out of and into the page.)

This diagram is correct, but it becomes less useful as the source charge distribution becomes more complicated. For example, consider the vector field diagram of a dipole ([link]).



The vector field of a dipole. Even with just two identical charges, the vector field diagram becomes difficult to understand.

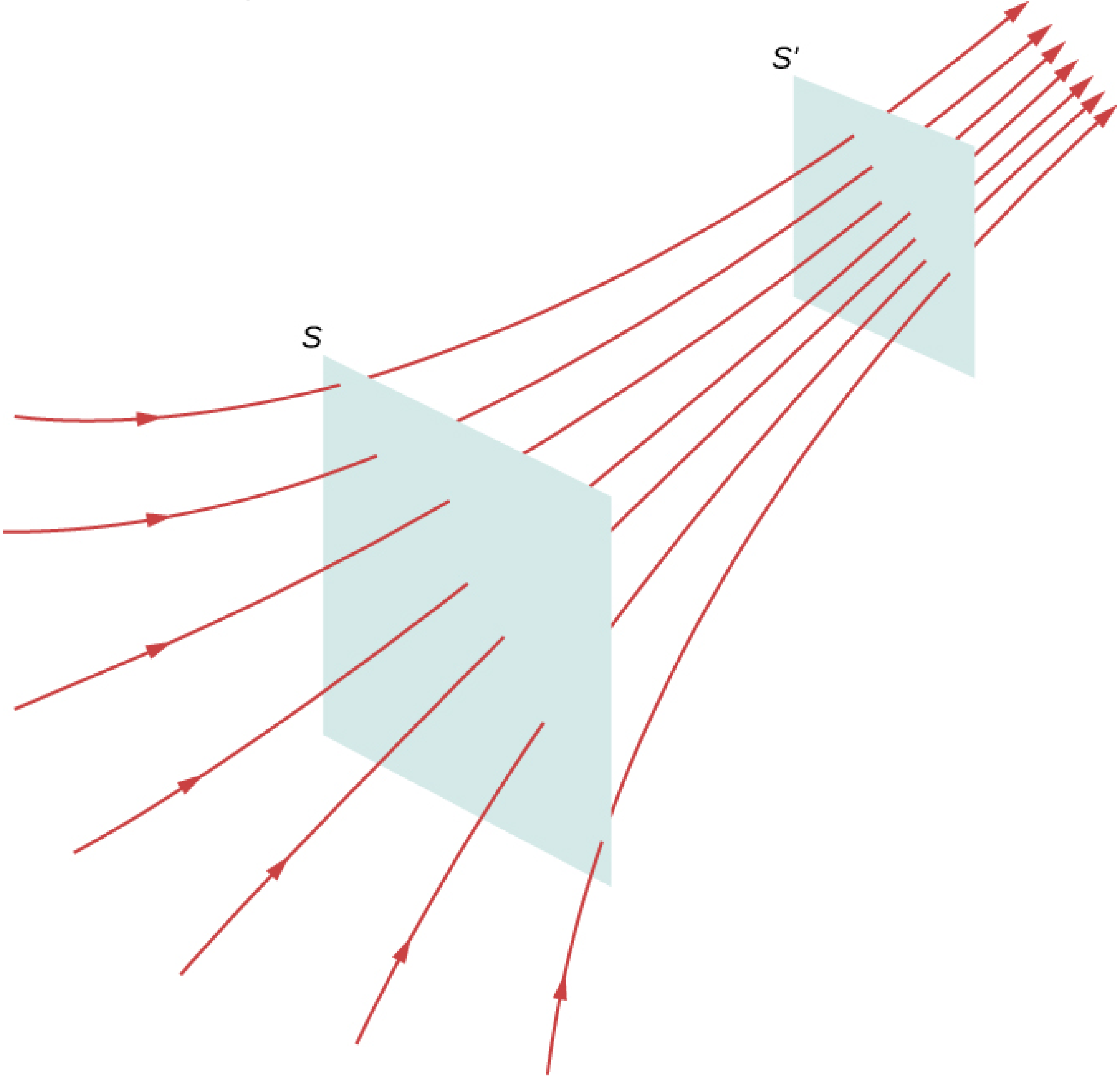
There is a more useful way to present the same information. Rather than drawing a large number of increasingly smaller vector arrows, we instead connect all of them together, forming continuous lines and curves, as shown in [link].



(a) The electric field line diagram of a positive point charge. (b) The field line diagram of a dipole. In both diagrams, the magnitude of the field is indicated by the field line density. The field *vectors* (not shown here) are everywhere tangent to the field lines.

Although it may not be obvious at first glance, these field diagrams convey the same information about the electric field as do the vector diagrams. First, the direction of the field at every point is simply the direction of the field vector at that same point. In other words, at any point in space, the field vector at each point is tangent to the field line at that same point. The arrowhead placed on a field line indicates its direction.

As for the magnitude of the field, that is indicated by the **field line density**—that is, the number of field lines per unit area passing through a small cross-sectional area perpendicular to the electric field. This field line density is drawn to be proportional to the magnitude of the field at that cross-section. As a result, if the field lines are close together (that is, the field line density is greater), this indicates that the magnitude of the field is large at that point. If the field lines are far apart at the cross-section, this indicates the magnitude of the field is small. [link] shows the idea.



Electric field lines passing through imaginary areas. Since the number of lines passing through each area is the same, but the areas themselves are different, the field line density is different. This indicates different magnitudes of the electric field at these points.

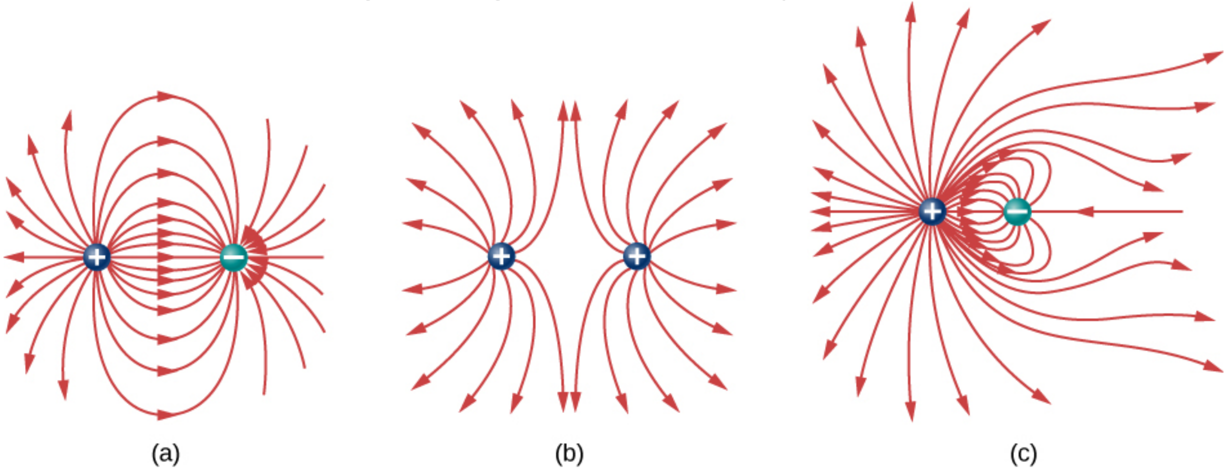
In [link], the same number of field lines passes through both surfaces (S and S'), but the surface S is larger than surface S'. Therefore, the density of field lines (number of lines per unit area) is larger at the location of S', indicating that the electric field is stronger at the location of S' than at S. The rules for creating an electric field diagram are as follows.

Note:
Problem-Solving Strategy: Drawing Electric Field Lines

1. Electric field lines either originate on positive charges or come in from infinity, and either terminate on negative charges or extend out to infinity.
2. The number of field lines originating or terminating at a charge is proportional to the magnitude of that charge. A charge of $2q$ will have twice as many lines as a charge of q.
3. At every point in space, the field vector at that point is tangent to the field line at that same point.
4. The field line density at any point in space is proportional to (and therefore is representative of) the magnitude of the field at that point in space.
5. Field lines can never cross. Since a field line represents the direction of the field at a given point, if two field lines crossed at some point, that would imply that the electric field was pointing in two different directions at a single point. This in turn would suggest that the (net) force on a test charge placed at that point would point in two different directions. Since this is obviously impossible, it follows that field lines must never cross.

Always keep in mind that field lines serve only as a convenient way to visualize the electric field; they are not physical entities. Although the direction and relative intensity of the electric field can be deduced from a set of field lines, the lines can also be misleading. For example, the field lines drawn to represent the electric field in a region must, by necessity, be discrete. However, the actual electric field in that region exists at every point in space.

Field lines for three groups of discrete charges are shown in [link]. Since the charges in parts (a) and (b) have the same magnitude, the same number of field lines are shown starting from or terminating on each charge. In (c), however, we draw three times as many field lines leaving the $+3q$ charge as entering the $-q$. The field lines that do not terminate at $-q$ emanate outward from the charge configuration, to infinity.



Three typical electric field diagrams. (a) A dipole. (b) Two identical charges. (c) Two charges with opposite signs and different magnitudes. Can you tell from the diagram which charge has the larger magnitude?

The ability to construct an accurate electric field diagram is an important, useful skill; it makes it much easier to estimate, predict, and therefore calculate the electric field of a source charge. The best way to develop this skill is with software that allows you to place source charges and then will draw the net field upon request. We strongly urge you to search the Internet for a program. Once you’ve found one you like, run several simulations to get the essential ideas of field diagram construction. Then practice drawing field diagrams, and checking your predictions with the computer-drawn diagrams.

Note:
One example of a field-line drawing program is from the PhET “Charges and Fields” simulation.

Summary

- Electric field diagrams assist in visualizing the field of a source charge.
- The magnitude of the field is proportional to the field line density.
- Field vectors are everywhere tangent to field lines.

Conceptual Questions

Exercise:

Problem:

If a point charge is released from rest in a uniform electric field, will it follow a field line? Will it do so if the electric field is not uniform?

Solution:

yes; no

Exercise:

Problem:

Under what conditions, if any, will the trajectory of a charged particle not follow a field line?

Exercise:

Problem:

How would you experimentally distinguish an electric field from a gravitational field?

Solution:

At the surface of Earth, the gravitational field is always directed in toward Earth’s center. An electric field could move a charged particle in a different direction than toward the center of Earth. This would indicate an electric field is present.

Exercise:

Problem:

A representation of an electric field shows 10 field lines perpendicular to a square plate. How many field lines should pass perpendicularly through the plate to depict a field with twice the magnitude?

Exercise:

Problem:

What is the ratio of the number of electric field lines leaving a charge 10*q* and a charge *q*?

Solution:

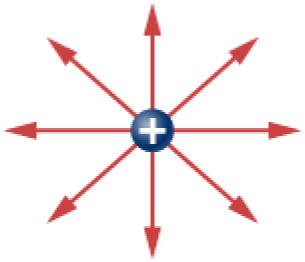
10

Problems

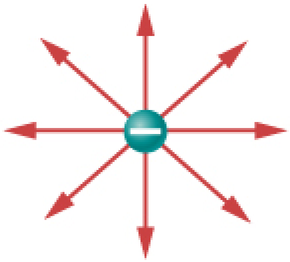
Exercise:

Problem:

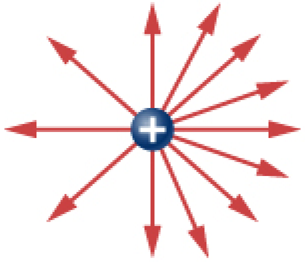
Which of the following electric field lines are incorrect for point charges? Explain why.



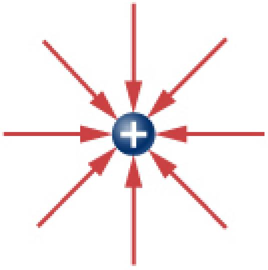
(a)



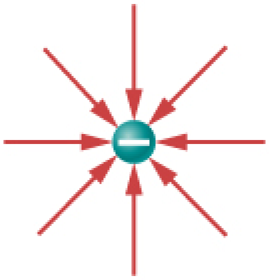
(b)



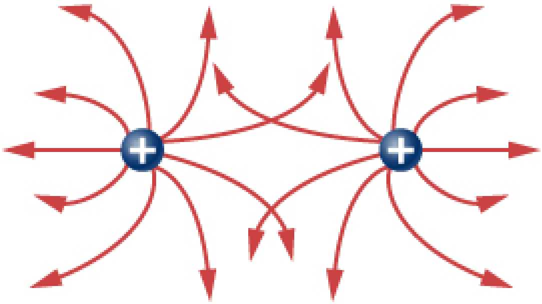
(c)



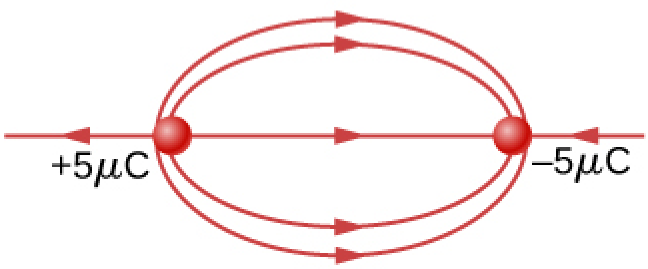
(d)



(e)



(f)



(g)

Exercise:

Problem:

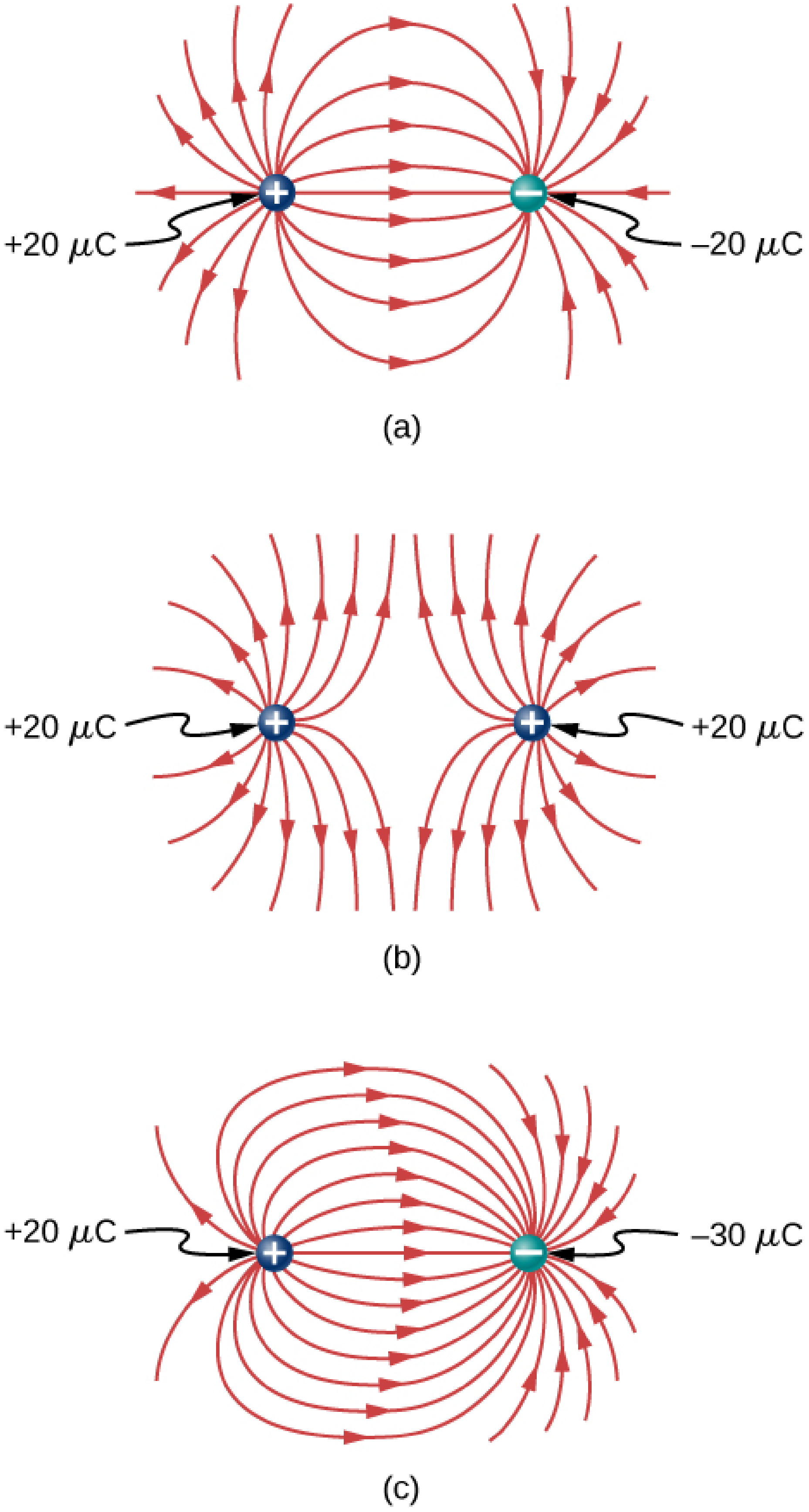
In this exercise, you will practice drawing electric field lines. Make sure you represent both the magnitude and direction of the electric field adequately. Note that the number of lines into or out of charges is proportional to the charges.

(a) Draw the electric field lines map for two charges $+20\ \mu\text{C}$ and $-20\ \mu\text{C}$ situated 5 cm from each other.

(b) Draw the electric field lines map for two charges $+20\ \mu\text{C}$ and $+20\ \mu\text{C}$ situated 5 cm from each other.

(c) Draw the electric field lines map for two charges $+20\ \mu\text{C}$ and $-30\ \mu\text{C}$ situated 5 cm from each other.

Solution:


+20 μC
–20 μC
(a)
+20 μC
+20 μC
(b)
+20 μC
–30 μC
(c)

Exercise:

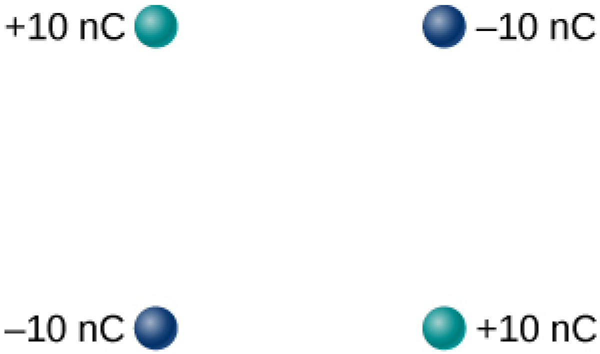
Problem:

Draw the electric field for a system of three particles of charges $+1\ \mu\text{C}$, $+2\ \mu\text{C}$, and $-3\ \mu\text{C}$ fixed at the corners of an equilateral triangle of side 2 cm.

Exercise:

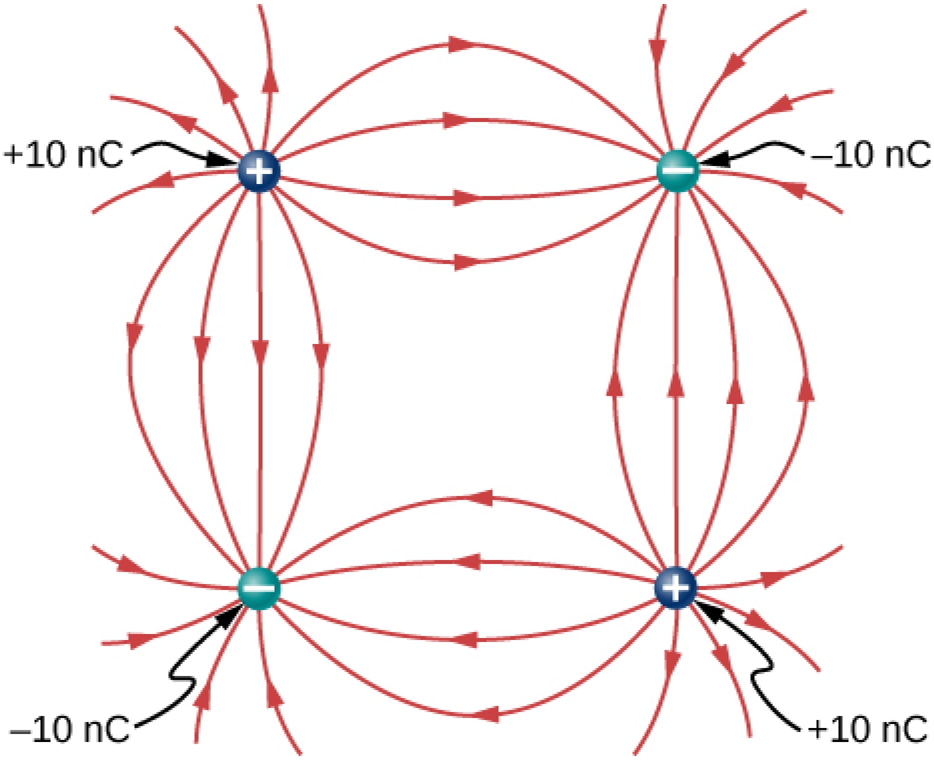
Problem:

Two charges of equal magnitude but opposite sign make up an electric dipole. A quadrupole consists of two electric dipoles that are placed anti-parallel at two edges of a square as shown.



Draw the electric field of the charge distribution.

Solution:



Exercise:

Problem:

Suppose the electric field of an isolated point charge decreased with distance as $1/r^{2+\delta}$ rather than as $1/r^2$. Show that it is then impossible to draw continous field lines so that their number per unit area is proportional to E.

Glossary

field line	smooth, usually curved line that indicates the direction of the electric field
field line density	number of field lines per square meter passing through an imaginary area; its purpose is to indicate the field strength at different points in space

Electric Dipoles

By the end of this section, you will be able to:

- Describe a permanent dipole
- Describe an induced dipole
- Define and calculate an electric dipole moment
- Explain the physical meaning of the dipole moment

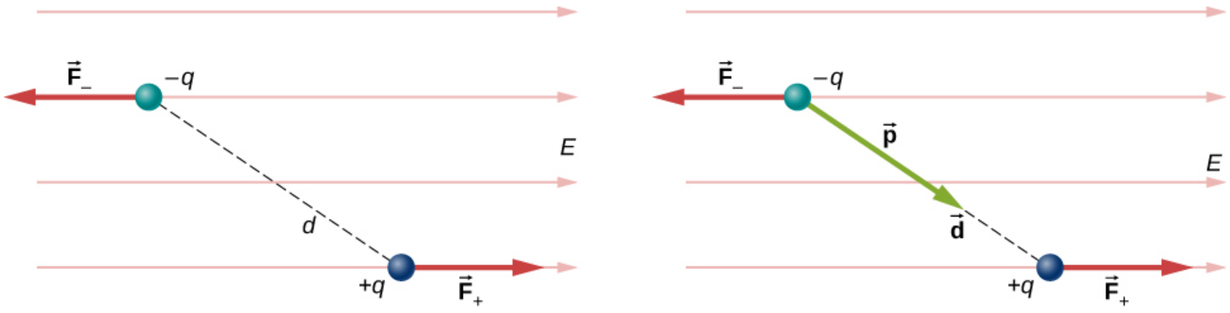
Earlier we discussed, and calculated, the electric field of a dipole: two equal and opposite charges that are “close” to each other. (In this context, “close” means that the distance *d* between the two charges is much, much less than the distance of the field point *P*, the location where you are calculating the field.) Let’s now consider what happens to a dipole when it is placed in an external field $\vec{\mathbf{E}}$. We assume that the dipole is a **permanent dipole**; it exists without the field, and does not break apart in the external field.

Rotation of a Dipole due to an Electric Field

For now, we deal with only the simplest case: The external field is uniform in space. Suppose we have the situation depicted in [link], where we denote the distance between the charges as the vector $\vec{\mathbf{d}}$, pointing from the negative charge to the positive charge. The forces on the two charges are equal and opposite, so there is no net force on the dipole. However, there is a torque:

Equation:

$$\begin{aligned}\vec{\tau} &= \left(\frac{\vec{\mathbf{d}}}{2} \times \vec{\mathbf{F}}_{+}\right) + \left(-\frac{\vec{\mathbf{d}}}{2} \times \vec{\mathbf{F}}_{-}\right) \\ &= \left[\left(\frac{\vec{\mathbf{d}}}{2}\right) \times \left(+q\vec{\mathbf{E}}\right) + \left(-\frac{\vec{\mathbf{d}}}{2}\right) \times \left(-q\vec{\mathbf{E}}\right)\right] \\ &= q\vec{\mathbf{d}} \times \vec{\mathbf{E}}.\end{aligned}$$



A dipole in an external electric field. (a) The net force on the dipole is zero, but the net torque is not. As a result, the dipole rotates, becoming aligned with the external field. (b) The dipole moment is a convenient way to characterize this effect. The $\vec{\mathbf{d}}$ points in the same direction as $\vec{\mathbf{p}}$.

The quantity $q\vec{\mathbf{d}}$ (the magnitude of each charge multiplied by the vector distance between them) is a property of the dipole; its value, as you can see, determines the torque that the dipole experiences in the external field. It is useful, therefore, to define this product as the so-called **dipole moment** of the dipole:

Note:
Equation:

$$\vec{\mathbf{p}} \equiv q\vec{\mathbf{d}}.$$

We can therefore write

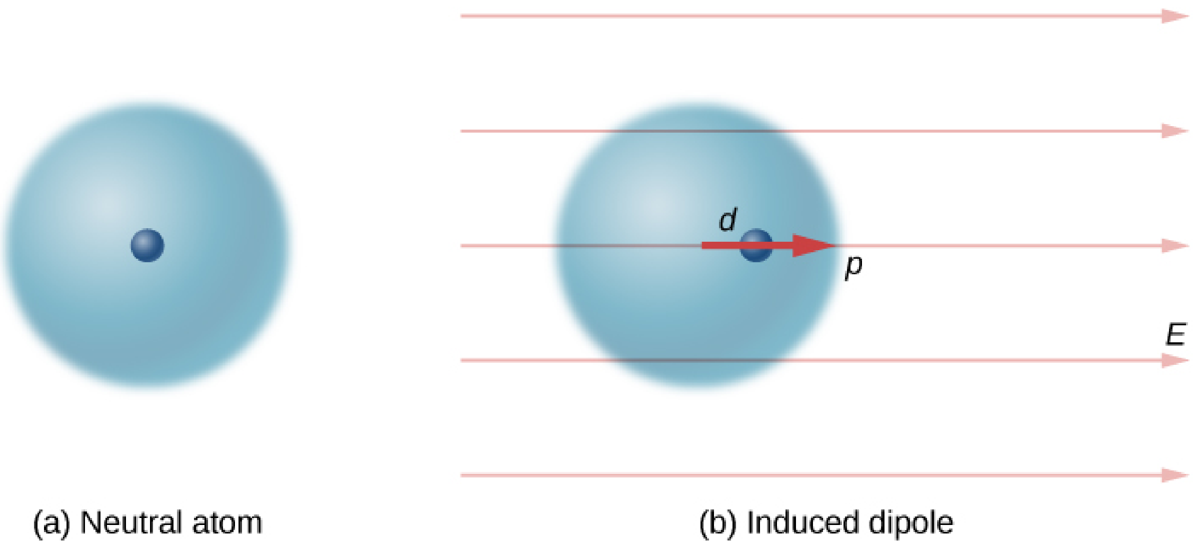
Note:
Equation:

$$\vec{\tau} = \vec{\mathbf{p}} \times \vec{\mathbf{E}}.$$

Recall that a torque changes the angular velocity of an object, the dipole, in this case. In this situation, the effect is to rotate the dipole (that is, align the direction of $\vec{\mathbf{p}}$) so that it is parallel to the direction of the external field.

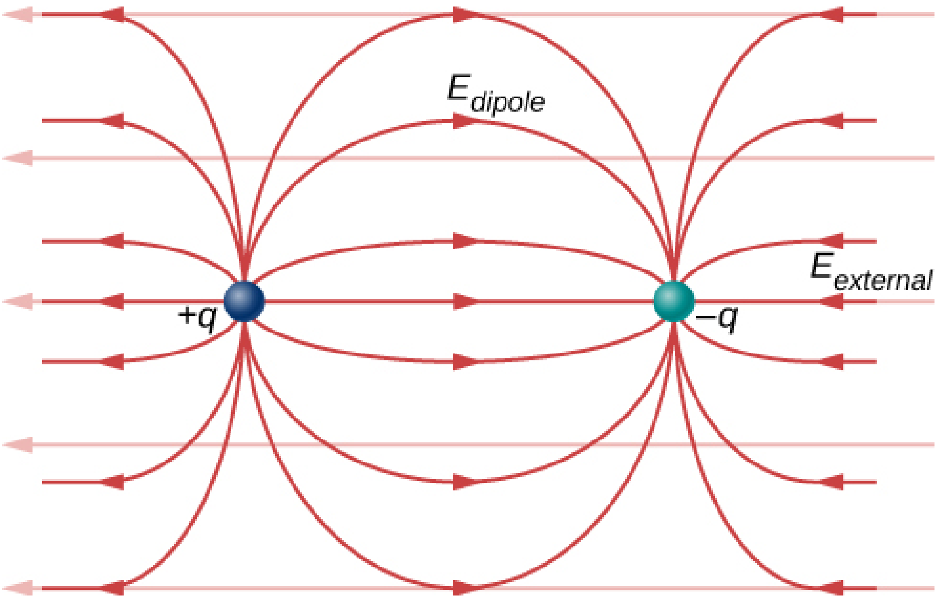
Induced Dipoles

Neutral atoms are, by definition, electrically neutral; they have equal amounts of positive and negative charge. Furthermore, since they are spherically symmetrical, they do not have a “built-in” dipole moment the way most asymmetrical molecules do. They obtain one, however, when placed in an external electric field, because the external field causes oppositely directed forces on the positive nucleus of the atom versus the negative electrons that surround the nucleus. The result is a new charge distribution of the atom, and therefore, an **induced dipole** moment ([link]).



A dipole is induced in a neutral atom by an external electric field. The induced dipole moment is aligned with the external field.

An important fact here is that, just as for a rotated polar molecule, the result is that the dipole moment ends up aligned parallel to the external electric field. Generally, the magnitude of an induced dipole is much smaller than that of an inherent dipole. For both kinds of dipoles, notice that once the alignment of the dipole (rotated or induced) is complete, the net effect is to decrease the total electric field $\vec{\mathbf{E}}_{\text{total}} = \vec{\mathbf{E}}_{\text{external}} + \vec{\mathbf{E}}_{\text{dipole}}$ in the regions inside the dipole charges ([link]). By “inside” we mean in between the charges. This effect is crucial for capacitors, as you will see in Capacitance.



The net electric field is the vector sum of the field of the dipole plus the external field.

Recall that we found the electric field of a dipole in [link]. If we rewrite it in terms of the dipole moment we get:

Equation:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\frac{\vec{\mathbf{p}}}{z^3}.$$

The form of this field is shown in [link]. Notice that along the plane perpendicular to the axis of the dipole and midway between the charges, the direction of the electric field is opposite that of the dipole and gets weaker the further from the axis one goes. Similarly, on the axis of the dipole (but outside it), the field points in the same direction as the dipole, again getting weaker the further one gets from the charges.

Summary

- If a permanent dipole is placed in an external electric field, it results in a torque that aligns it with the external field.
- If a nonpolar atom (or molecule) is placed in an external field, it gains an induced dipole that is aligned with the external field.
- The net field is the vector sum of the external field plus the field of the dipole (physical or induced).
- The strength of the polarization is described by the dipole moment of the dipole, $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$.

Key Equations

Coulomb’s law	$\vec{\mathbf{F}}_{12}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r_{12}^2}\hat{\mathbf{r}}_{12}$
Superposition of electric forces	$\vec{\mathbf{F}}(r) = \frac{1}{4\pi\varepsilon_0}Q\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{\mathbf{r}}_i$
Electric force due to an electric field	$\vec{\mathbf{F}} = Q\vec{\mathbf{E}}$
Electric field at point *P*	$\vec{\mathbf{E}}(P) \equiv \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{\mathbf{r}}_i$
Field of an infinite wire	

	$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0}\ \frac{2\lambda}{z}\widehat{\mathbf{k}}$
Field of an infinite plane	$\vec{\mathbf{E}} = \frac{\sigma}{2\varepsilon_0}\widehat{\mathbf{k}}$
Dipole moment	$\vec{\mathbf{p}} \equiv q\vec{\mathbf{d}}$
Torque on dipole in external E-field	$\vec{\tau} = \vec{\mathbf{p}}\ \times\ \vec{\mathbf{E}}$

Conceptual Questions

Exercise:

Problem:

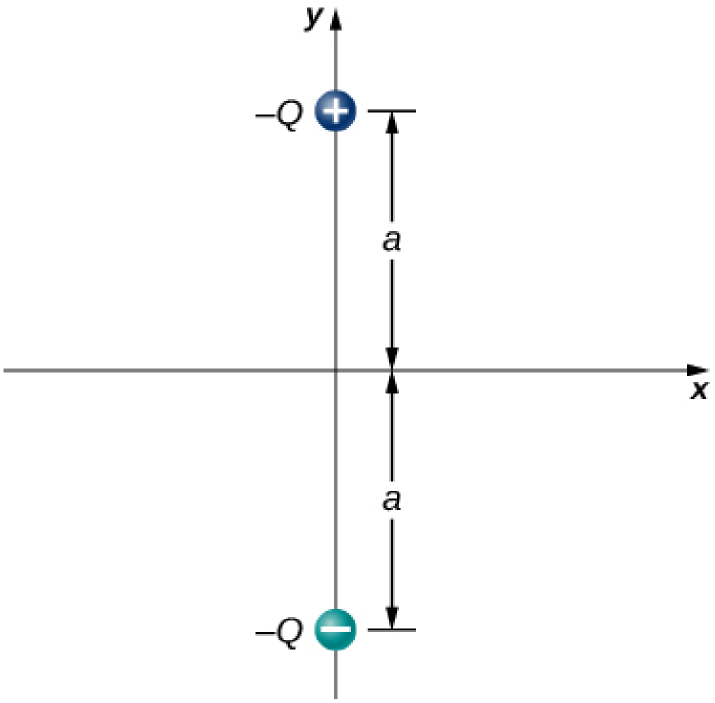
What are the stable orientation(s) for a dipole in an external electric field? What happens if the dipole is slightly perturbed from these orientations?

Problems

Exercise:

Problem:

Consider the equal and opposite charges shown below. (a) Show that at all points on the *x*-axis for which $|x| \gg a, E \approx Qa/2\pi\varepsilon_0 x^3$. (b) Show that at all points on the *y*-axis for which $|y| \gg a, E \approx Qa/\pi\varepsilon_0 y^3$.



Solution:

$$
\begin{aligned}
&E_x = 0,\\
&E_y = \frac{1}{4\pi\varepsilon_0}\left[\frac{2q}{(x^2+a^2)}\ \frac{a}{\sqrt{(x^2+a^2)}}\right]\\
&\Rightarrow x \gg a \Rightarrow \frac{1}{2\pi\varepsilon_0}\ \frac{qa}{x^3},\\
&E_y = \frac{q}{4\pi\varepsilon_0}\left[\frac{2ya+2ya}{(y-a)^2(y+a)^2}\right]\\
&\Rightarrow y \gg a \Rightarrow \frac{1}{\pi\varepsilon_0}\ \frac{qa}{y^3}
\end{aligned}
$$

Exercise:

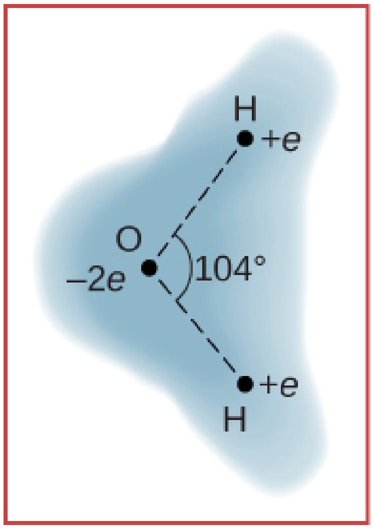
Problem:

(a) What is the dipole moment of the configuration shown above? If $Q = 4.0\,\mu\text{C}$, (b) what is the torque on this dipole with an electric field of $4.0\ \times\ 10^5\ \text{N/C}\hat{\mathbf{i}}$? (c) What is the torque on this dipole with an electric field of $-4.0\ \times\ 10^5\ \text{N/C}\hat{\mathbf{i}}$? (d) What is the torque on this dipole with an electric field of $\pm 4.0\ \times\ 10^5\ \text{N/C}\hat{\mathbf{j}}$?

Exercise:

Problem:

A water molecule consists of two hydrogen atoms bonded with one oxygen atom. The bond angle between the two hydrogen atoms is 104° (see below). Calculate the net dipole moment of a hypothetical water molecule where the charge at the oxygen molecule is −2e and at each hydrogen atom is +e. The net dipole moment of the molecule is the vector sum of the individual dipole moment between the two O-Hs. The separation O-H is 0.9578 angstroms.



Solution:

The net dipole moment of the molecule is the vector sum of the individual dipole moments between the two O-H. The separation O-H is 0.9578 angstroms:

$\vec{\mathbf{p}} = 1.889 \ \times \ 10^{-29} \ \text{Cm} \ \hat{\mathbf{i}}$

Additional Problems

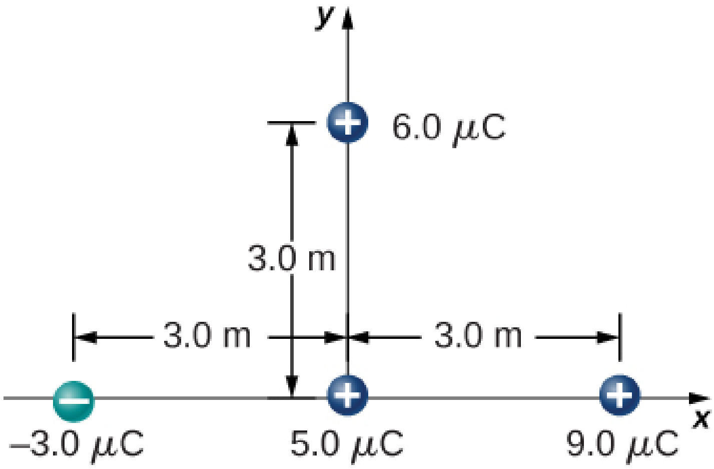
Exercise:

Problem:

Point charges $q_1 = 2.0\ \mu\text{C}$ and $q_1 = 4.0\ \mu\text{C}$ are located at $r_1 = \left(4.0\hat{\mathbf{i}} - 2.0\hat{\mathbf{j}} + 2.0\widehat{\mathbf{k}}\right)\text{m}$ and $r_2 = \left(8.0\hat{\mathbf{i}} + 5.0\hat{\mathbf{j}} - 9.0\widehat{\mathbf{k}}\right)\text{m}$. What is the force of q_2 on q_1?

Exercise:

Problem: What is the force on the 5.0-μC charge shown below?



Solution:

$$\vec{\mathbf{F}}_{\text{net}} = [-8.99 \times 10^9 \frac{3.0 \times 10^{-6}(5.0 \times 10^{-6})}{(3.0\,\text{m})^2} - 8.99 \times 10^9 \frac{9.0 \times 10^{-6}(5.0 \times 10^{-6})}{(3.0\,\text{m})^2}]\hat{\mathbf{i}}$$

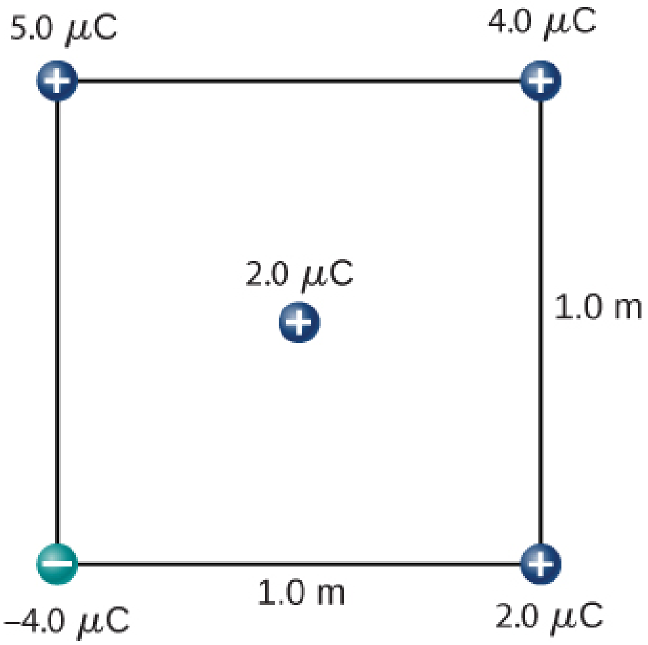
,

$$-8.99 \times 10^9 \frac{6.0 \times 10^{-6}(5.0 \times 10^{-6})}{(3.0\,\text{m})^2}\hat{\mathbf{j}} = -0.06\ \text{N}\hat{\mathbf{i}} - 0.03\ \text{N}\hat{\mathbf{j}}$$

Exercise:

Problem:

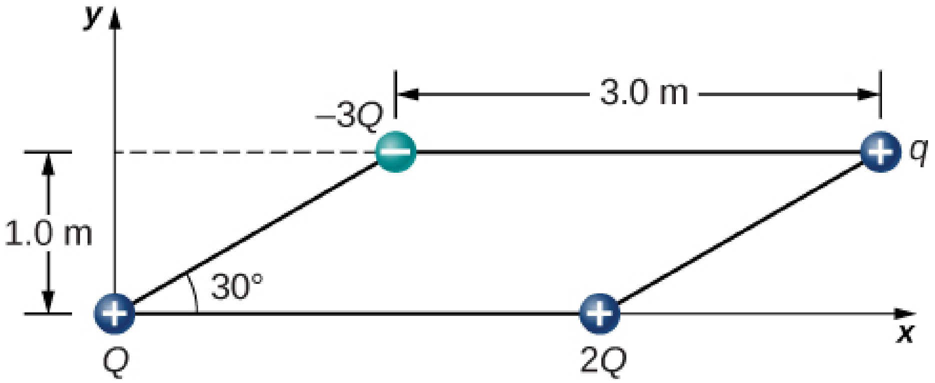
What is the force on the 2.0-μC charge placed at the center of the square shown below?


5.0 μC
4.0 μC
2.0 μC
1.0 m
1.0 m
–4.0 μC
2.0 μC

Exercise:

Problem:

Four charged particles are positioned at the corners of a parallelogram as shown below. If $q = 5.0\ \mu\text{C}$ and $Q = 8.0\ \mu\text{C}$, what is the net force on q?


y
–3Q
3.0 m
q
1.0 m
30°
x
Q
2Q

Solution:

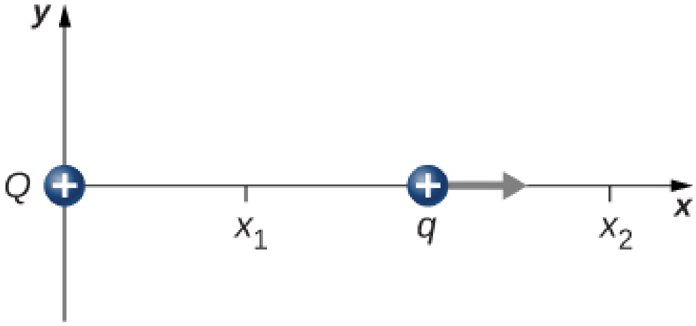
Charges Q and q form a right triangle of sides 1 m and $3 + \sqrt{3}$ m.
Charges $2Q$ and q form a right triangle of sides 1 m and $\sqrt{3}$ m.
$F_x \quad = 0.049\,\text{N},$

$F_y = 0.093 \text{ N},$
$\vec{\mathbf{F}}_{\text{net}} = 0.036 \text{ N}\,\hat{\mathbf{i}} + 0.09 \text{ N}\,\hat{\mathbf{j}}$

Exercise:

Problem:

A charge Q is fixed at the origin and a second charge q moves along the *x*-axis, as shown below. How much work is done on q by the electric force when q moves from x_1 to x_2?



Exercise:

Problem:

A charge $q = -2.0\ \mu\text{C}$ is released from rest when it is 2.0 m from a fixed charge $Q = 6.0\ \mu\text{C}$. What is the kinetic energy of q when it is 1.0 m from Q?

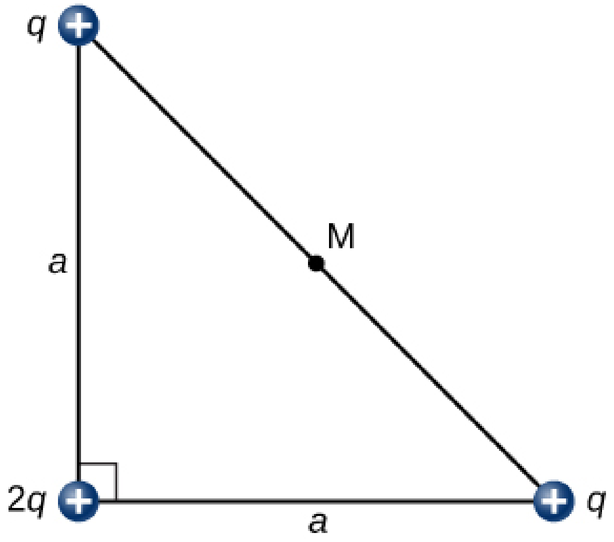
Solution:

$W = 0.054 \text{ J}$

Exercise:

Problem:

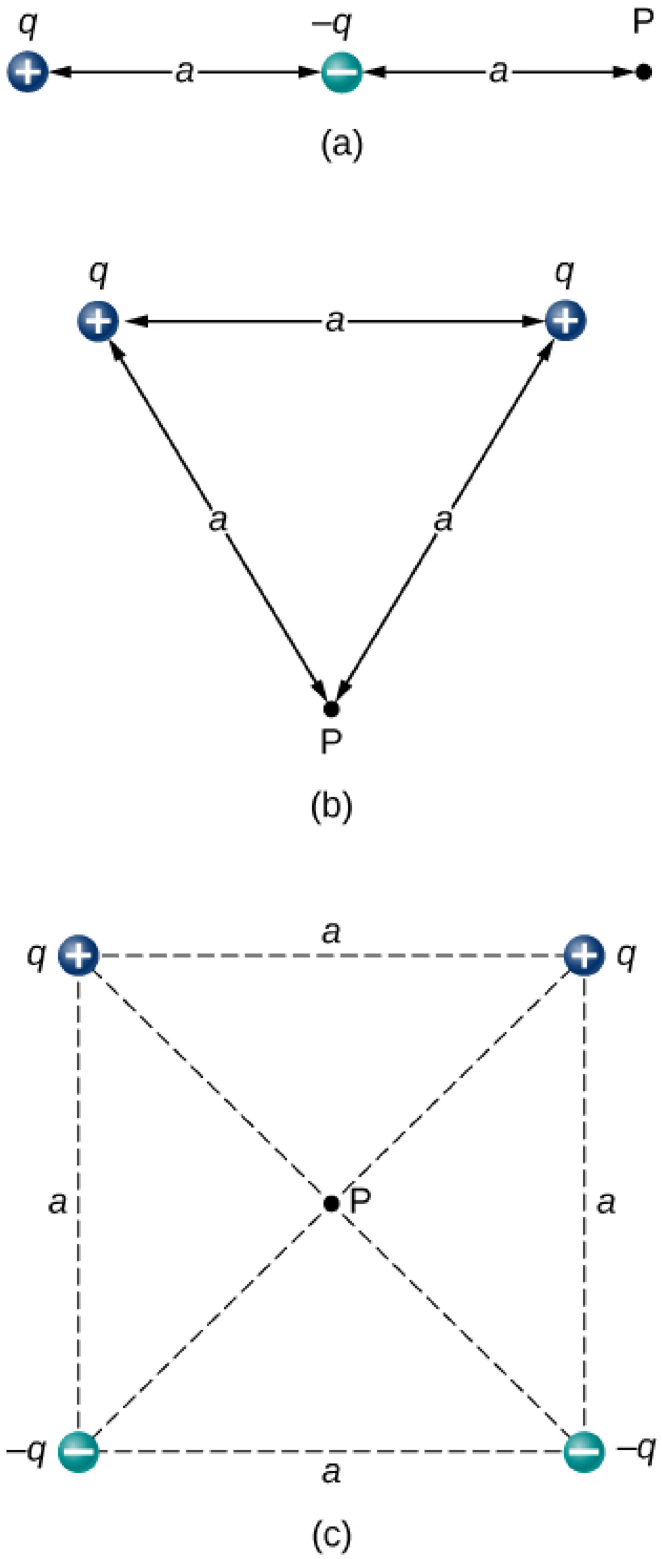
What is the electric field at the midpoint M of the hypotenuse of the triangle shown below?



Exercise:

Problem:

Find the electric field at *P* for the charge configurations shown below.


q
−q
P
a
a
q
q
a
a
a
P
q
q
a
a
a
P
a
−q
−q

(a)

(b)

(c)

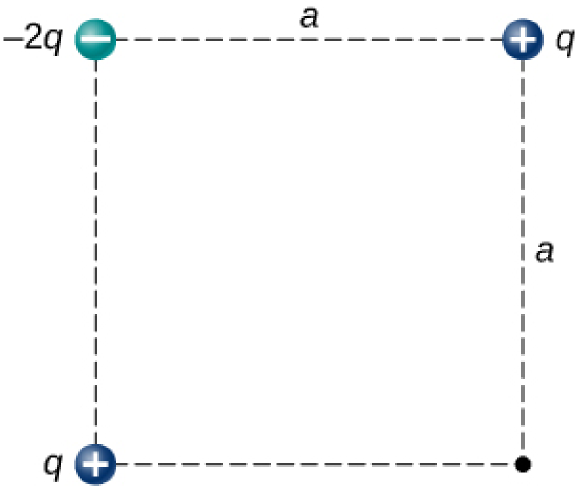
Solution:

a. $\vec{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{(2a)^2} - \frac{q}{a^2}\right)\hat{\mathbf{i}}$; b. $\vec{\mathbf{E}} = \frac{\sqrt{3}}{4\pi\varepsilon_0}\frac{q}{a^2}(-\hat{\mathbf{j}})$; c. $\vec{\mathbf{E}} = \frac{2}{\pi\varepsilon_0}\frac{q}{a^2}\frac{1}{\sqrt{2}}(-\hat{\mathbf{j}})$

Exercise:

Problem:

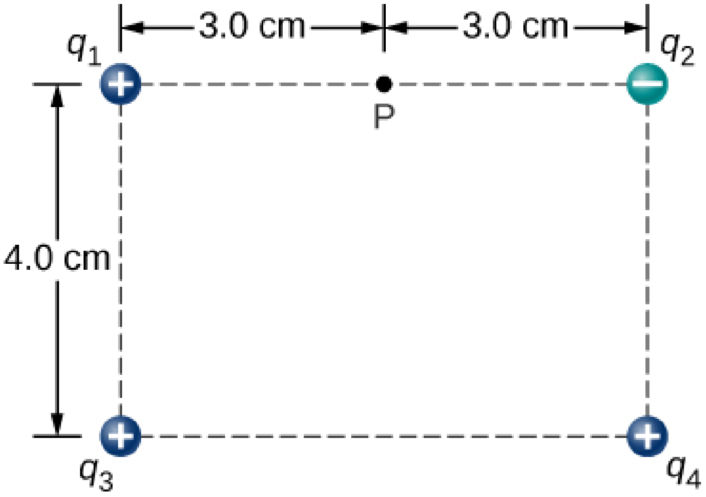
(a) What is the electric field at the lower-right-hand corner of the square shown below? (b) What is the force on a charge *q* placed at that point?



Exercise:

Problem:

Point charges are placed at the four corners of a rectangle as shown below: $q_1 = 2.0 \times 10^{-6}$ C, $q_2 = -2.0 \times 10^{-6}$ C, $q_3 = 4.0 \times 10^{-6}$ C, and $q_4 = 1.0 \times 10^{-6}$ C. What is the electric field at *P*?


3.0 cm
3.0 cm
q_1
q_2
P
4.0 cm
q_3
q_4

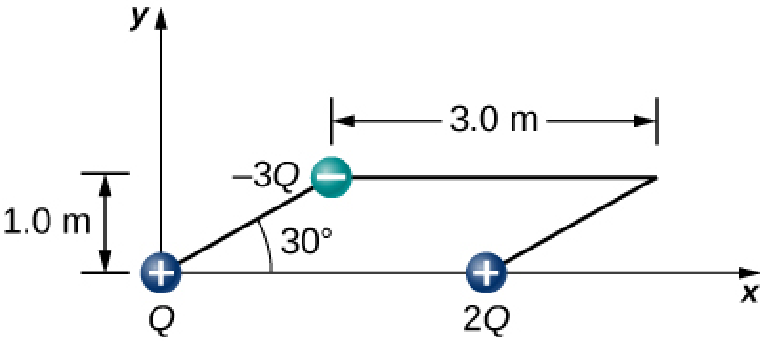
Solution:

$$\vec{\mathbf{E}}_{\text{net}} = \vec{\mathbf{E}}_1 + \vec{\mathbf{E}}_2 + \vec{\mathbf{E}}_3 + \vec{\mathbf{E}}_4 = \left(4.65\hat{\mathbf{i}} + 1.44\hat{\mathbf{j}}\right) \times 10^7 \text{ N/C}$$

Exercise:

Problem:

Three charges are positioned at the corners of a parallelogram as shown below. (a) If $Q = 8.0\ \mu\text{C}$, what is the electric field at the unoccupied corner? (b) What is the force on a 5.0-μC charge placed at this corner?


y
3.0 m
−3Q
1.0 m
30°
Q
2Q
x

Exercise:

Problem:

A positive charge q is released from rest at the origin of a rectangular coordinate system and moves under the influence of the electric field $\vec{\mathbf{E}} = E_0\,(1 + x/a)\hat{\mathbf{i}}$. What is the kinetic energy of q when it passes through $x = 3a$?

Solution:

$F = qE_0\,(1 + x/a)\;\; W = \frac{1}{2}m(v^2 - v_0^2),$
$\frac{1}{2}mv^2 = qE_0(\frac{15a}{2})\,\mathrm{J}$

Exercise:

Problem:

A particle of charge $-q$ and mass *m* is placed at the center of a uniformaly charged ring of total charge *Q* and radius *R*. The particle is displaced a small distance along the axis perpendicular to the plane of the ring and released. Assuming that the particle is constrained to move along the axis, show that the particle oscillates in simple harmonic motion with a frequency $f = \frac{1}{2\pi}\sqrt{\frac{qQ}{4\pi\varepsilon_0 mR^3}}$.

Exercise:

Problem:

Charge is distributed uniformly along the entire *y*-axis with a density λ_y and along the positive *x*-axis from $x = a$ to $x = b$ with a density λ_x. What is the force between the two distributions?

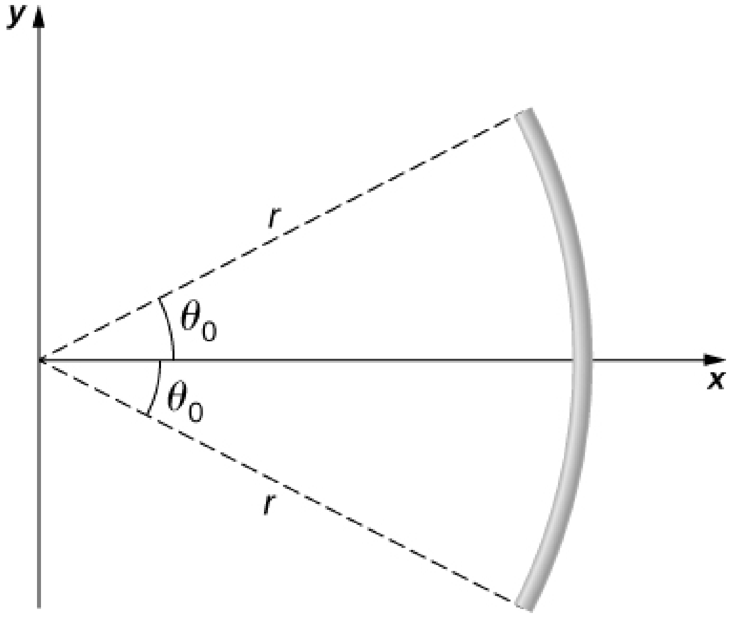
Solution:

Electric field of wire at *x*: $\vec{\mathbf{E}}(x) = \frac{1}{4\pi\varepsilon_0}\,\frac{2\lambda_y}{x}\hat{\mathbf{i}},$
$dF = \frac{\lambda_y\lambda_x}{2\pi\varepsilon_0}(\ln b - \ln a)$

Exercise:

Problem:

The circular arc shown below carries a charge per unit length $\lambda = \lambda_0 \cos\theta$, where θ is measured from the *x*-axis. What is the electric field at the origin?

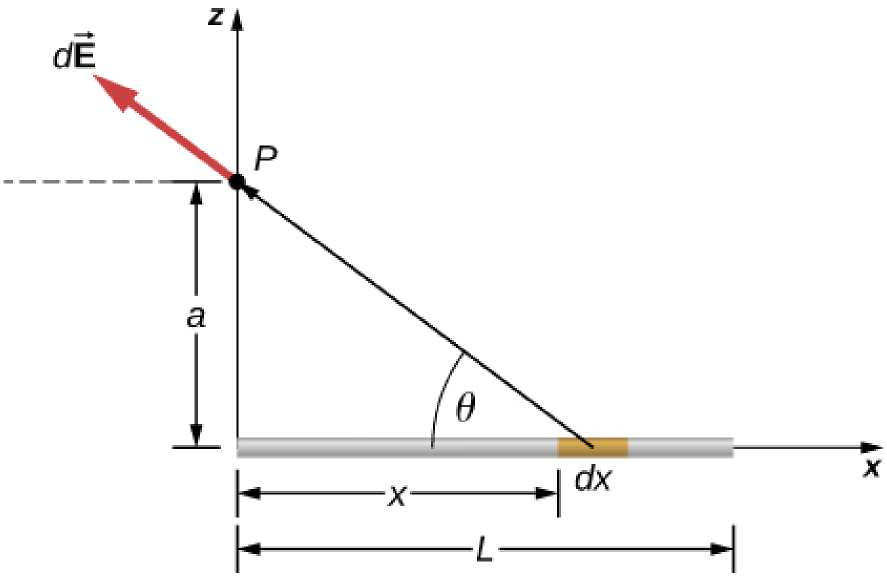


Exercise:

Problem:

Calculate the electric field due to a uniformly charged rod of length *L*, aligned with the *x*-axis with one end at the origin; at a point *P* on the *z*-axis.

Solution:



$$dE_x = \frac{1}{4\pi\varepsilon_0} \frac{\lambda dx}{(x^2+a^2)} \frac{x}{\sqrt{x^2+a^2}},$$

$$\vec{\mathbf{E}}_x = \frac{\lambda}{4\pi\varepsilon_0} \left[\frac{1}{\sqrt{L^2+a^2}} - \frac{1}{a} \right] \hat{\mathbf{i}},$$

$$dE_z = \frac{1}{4\pi\varepsilon_0} \frac{\lambda dx}{(x^2+a^2)} \frac{a}{\sqrt{x^2+a^2}},$$

$$\vec{\mathbf{E}}_z = \frac{\lambda}{4\pi\varepsilon_0 a} \frac{L}{\sqrt{L^2+a^2}} \widehat{\mathbf{k}},$$

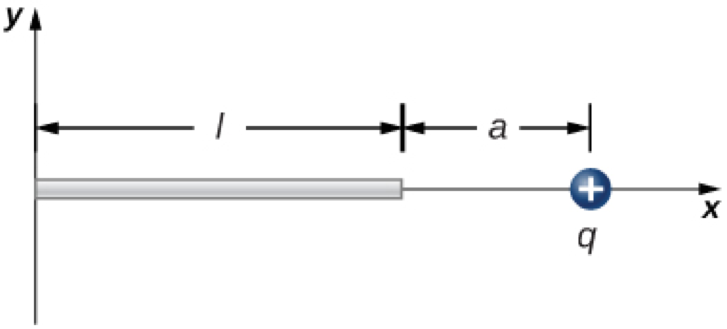
Substituting *z* for *a*, we have:

$$\vec{\mathbf{E}}(z) = \frac{\lambda}{4\pi\varepsilon_0} \left[\frac{1}{\sqrt{L^2+z^2}} - \frac{1}{z} \right] \hat{\mathbf{i}} + \frac{\lambda}{4\pi\varepsilon_0 z} \frac{L}{\sqrt{L^2+z^2}} \widehat{\mathbf{k}}$$

Exercise:

Problem:

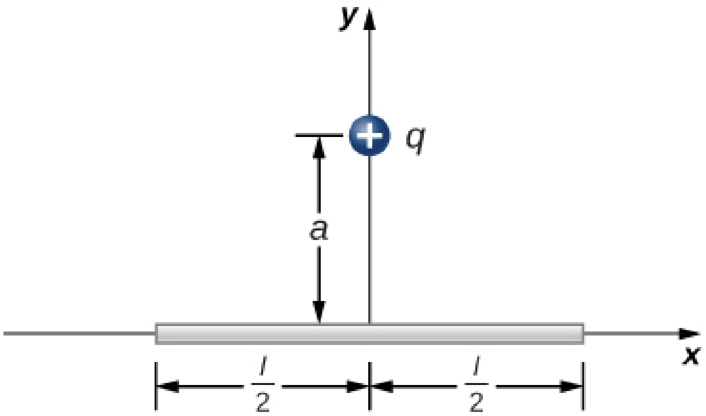
The charge per unit length on the thin rod shown below is λ. What is the electric force on the point charge *q*? Solve this problem by first considering the electric force $d\vec{\mathbf{F}}$ on *q* due to a small segment dx of the rod, which contains charge λdx. Then, find the net force by integrating $d\vec{\mathbf{F}}$ over the length of the rod.



Exercise:

Problem:

The charge per unit length on the thin rod shown here is λ. What is the electric force on the point charge *q*? (See the preceding problem.)



Solution:

There is a net force only in the *y*-direction. Let θ be the angle the vector from *dx* to *q* makes with the *x*-axis. The components along the *x*-axis cancel due to symmetry, leaving the *y*-component of the force.

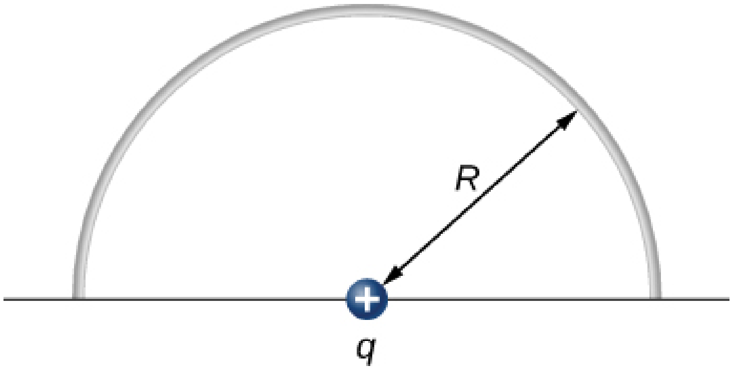
$$dF_{\mathrm{y}} = \frac{1}{4\pi\varepsilon_0}\frac{aq\lambda dx}{\left(x^2+a^2\right)^{3/2}},$$

$$F_y = \frac{1}{2\pi\varepsilon_0}\frac{q\lambda}{a}\left[\frac{l/2}{\left((l/2)^2+a^2\right)^{1/2}}\right]$$

Exercise:

Problem:

The charge per unit length on the thin semicircular wire shown below is λ. What is the electric force on the point charge q? (See the preceding problems.)



Glossary

dipole moment

property of a dipole; it characterizes the combination of distance between the opposite charges, and the magnitude of the charges

induced dipole

typically an atom, or a spherically symmetric molecule; a dipole created due to opposite forces displacing the positive and negative charges

permanent dipole

typically a molecule; a dipole created by the arrangement of the charged particles from which the dipole is created

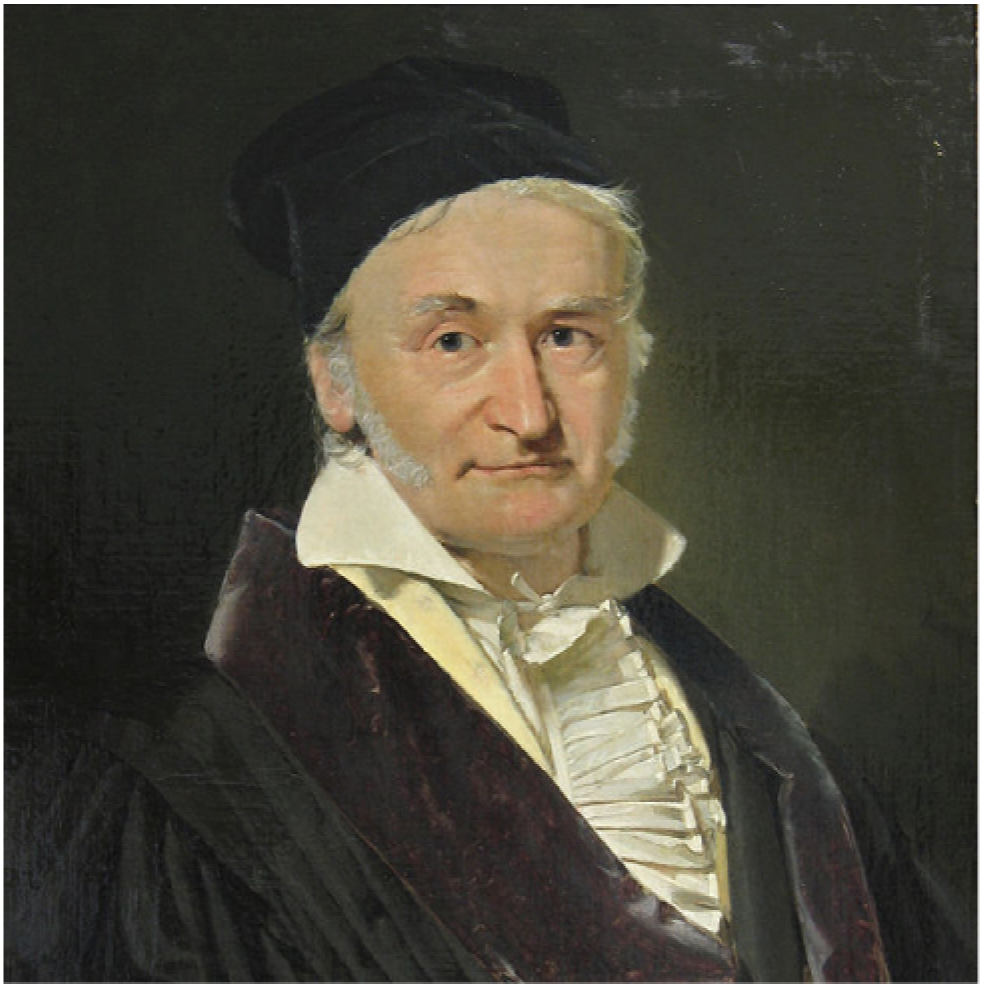
Introduction class="introduction"

This chapter introduces the concept of flux, which relates a physical quantity and the area through which it is flowing. Although we introduce this concept with the electric field, the concept may be used for many other quantities, such as fluid flow. (credit: modification of work by “Alessandro”/Flickr)

Flux is a general and broadly applicable concept in physics. However, in this chapter, we concentrate on the flux of the electric field. This allows us to introduce Gauss’s law, which is particularly useful for finding the electric fields of charge distributions exhibiting spatial symmetry. The main topics discussed here are

1. **Electric flux.** We define electric flux for both open and closed surfaces.
2. **Gauss’s law.** We derive Gauss’s law for an arbitrary charge distribution and examine the role of electric flux in Gauss’s law.
3. **Calculating electric fields with Gauss’s law.** The main focus of this chapter is to explain how to use Gauss’s law to find the electric fields of spatially symmetrical charge distributions. We discuss the importance of choosing a Gaussian surface and provide examples involving the applications of Gauss’s law.
4. **Electric fields in conductors.** Gauss’s law provides useful insight into the absence of electric fields in conducting materials.

So far, we have found that the electrostatic field begins and ends at point charges and that the field of a point charge varies inversely with the square of the distance from that charge. These characteristics of the electrostatic field lead to an important mathematical relationship known as Gauss’s law. This law is named in honor of the extraordinary German mathematician and scientist Karl Friedrich Gauss ([link]). Gauss’s law gives us an elegantly simple way of finding the electric field, and, as you will see, it can be much easier to use than the integration method described in the previous chapter. However, there is a catch—Gauss’s law has a limitation in that, while always true, it can be readily applied only for charge distributions with certain symmetries.



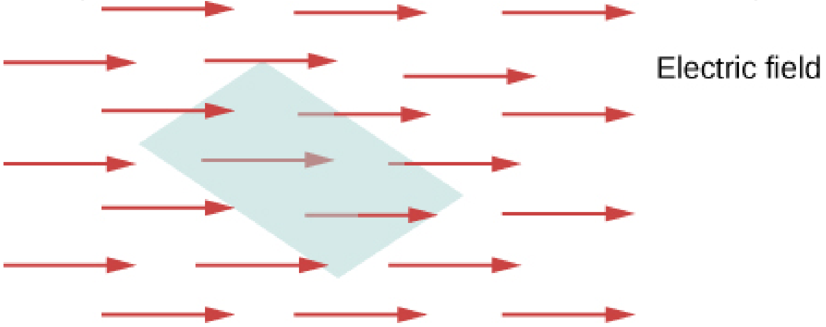
Karl Friedrich Gauss (1777–1855) was a legendary mathematician of the nineteenth century. Although his major contributions were to the field of mathematics, he also did important work in physics and astronomy.

Electric Flux

By the end of this section, you will be able to:

- Define the concept of flux
- Describe electric flux
- Calculate electric flux for a given situation

The concept of **flux** describes how much of something goes through a given area. More formally, it is the dot product of a vector field (in this chapter, the electric field) with an area. You may conceptualize the flux of an electric field as a measure of the number of electric field lines passing through an area ([link]). The larger the area, the more field lines go through it and, hence, the greater the flux; similarly, the stronger the electric field is (represented by a greater density of lines), the greater the flux. On the other hand, if the area rotated so that the plane is aligned with the field lines, none will pass through and there will be no flux.

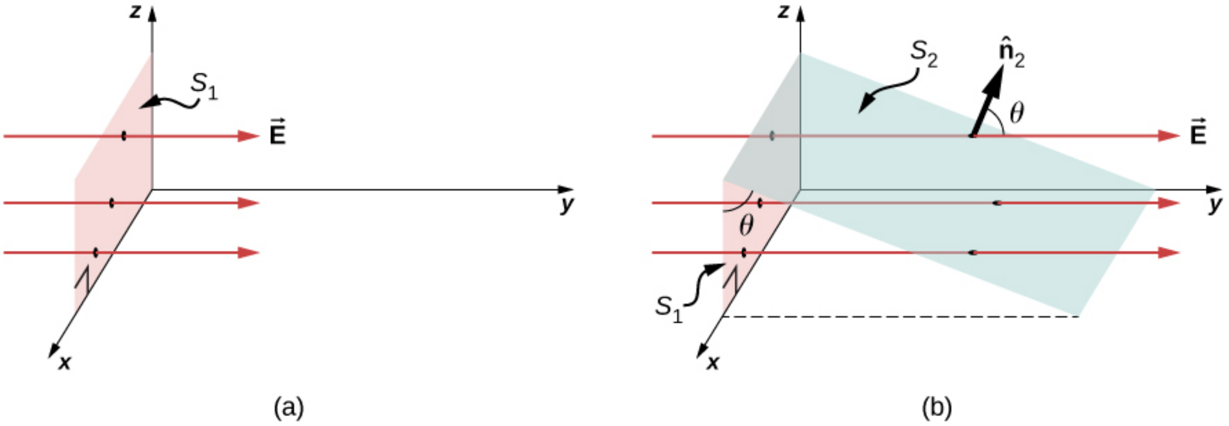


The flux of an electric field through the shaded area captures information about the “number” of electric field lines passing through the area. The numerical value of the electric flux depends on the magnitudes of the electric field and the area, as well as the relative orientation of the area with respect to the direction of the electric field.

A macroscopic analogy that might help you imagine this is to put a hula hoop in a flowing river. As you change the angle of the hoop relative to the direction of the current, more or less of the flow will go through the hoop. Similarly, the amount of flow through the hoop depends on the strength of the current and the size of the hoop. Again, flux is a general concept; we can also use it to describe the amount of sunlight hitting a solar panel or the amount of energy a telescope receives from a distant star, for example.

To quantify this idea, [link](a) shows a planar surface S_1 of area A_1 that is perpendicular to the uniform electric field $\overrightarrow{\mathbf{E}} = E\widehat{\mathbf{y}}$. If N field lines pass through S_1, then we know from the definition of electric field lines (Electric Charges and Fields) that $N/A_1 \propto E$, or $N \propto EA_1$.

The quantity EA_1 is the **electric flux** through S_1. We represent the electric flux through an open surface like S_1 by the symbol Φ. Electric flux is a scalar quantity and has an SI unit of newton-meters squared per coulomb ($\mathrm{N \cdot m^2/C}$). Notice that $N \propto EA_1$ may also be written as $N \propto \Phi$, demonstrating that *electric flux is a measure of the number of field lines crossing a surface*.



(a) A planar surface S_1 of area A_1 is perpendicular to the electric field $E\hat{\mathbf{j}}$. N field lines cross surface S_1. (b) A surface S_2 of area A_2 whose projection onto the *xz*-plane is S_1.The same number of field lines cross each surface.

Now consider a planar surface that is not perpendicular to the field. How would we represent the electric flux? [link](b) shows a surface S_2 of area A_2 that is inclined at an angle θ to the *xz*-plane and whose projection in that plane is S_1 (area A_1). The areas are related by $A_2 \cos\theta = A_1$. Because the same number of field lines crosses both S_1 and S_2, the fluxes through both surfaces must be the same. The flux through S_2 is therefore $\Phi = EA_1 = EA_2 \cos\theta$. Designating $\widehat{\mathbf{n}}_2$ as a unit vector normal to S_2 (see [link](b)), we obtain

Equation:

$$\Phi = \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}_2 A_2.$$

Note:
Check out this video to observe what happens to the flux as the area changes in size and angle, or the electric field changes in strength.

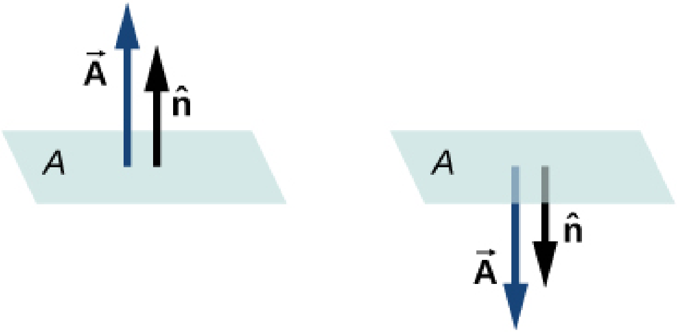
Area Vector

For discussing the flux of a vector field, it is helpful to introduce an area vector $\overrightarrow{\mathbf{A}}$. This allows us to write the last equation in a more compact form. What should the magnitude of the area vector be? What should the direction of the area vector be? What are the implications of how you answer the previous question?

The **area vector** of a flat surface of area A has the following magnitude and direction:

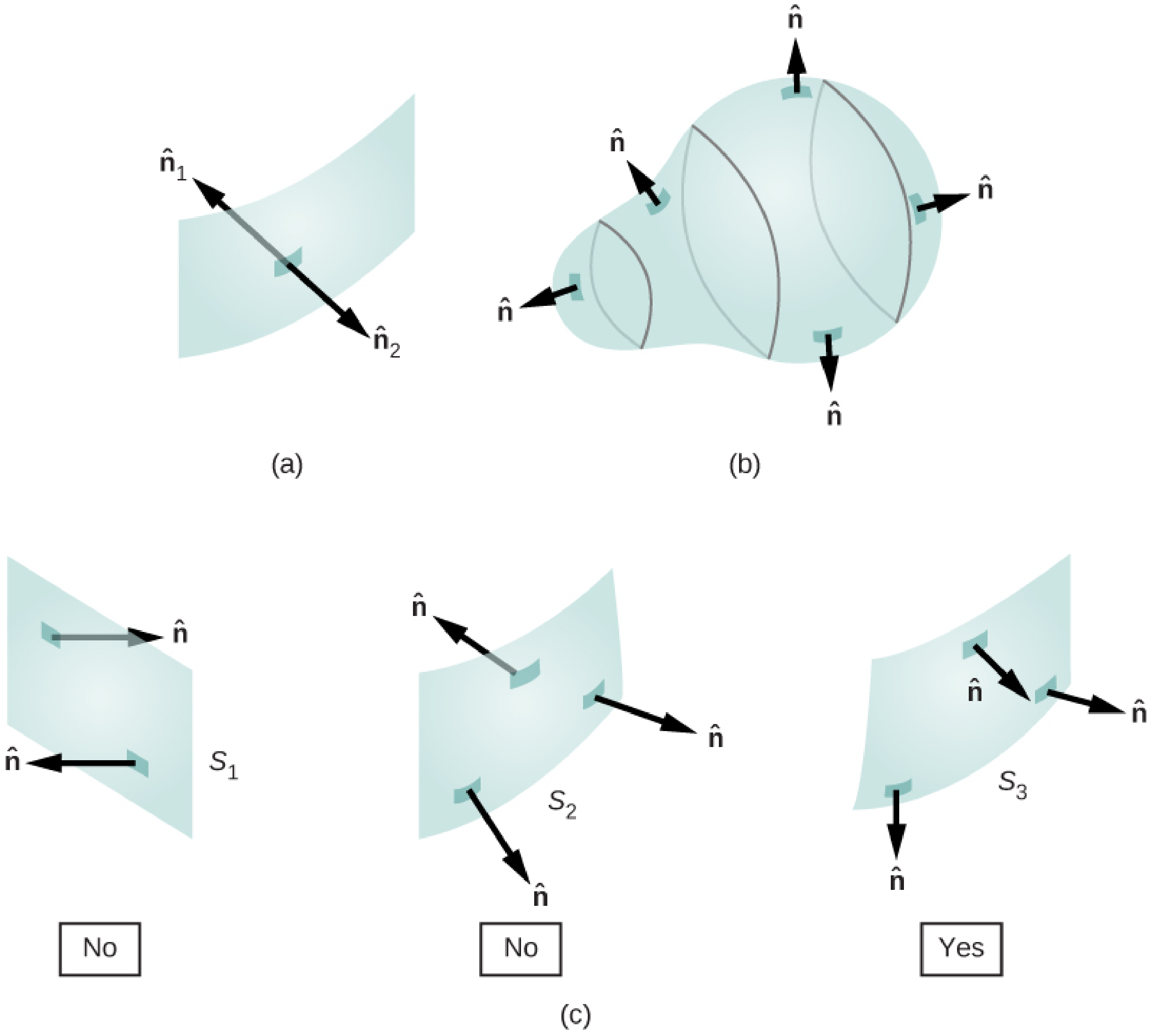
- Magnitude is equal to area (A)
- Direction is along the normal to the surface ($\hat{\mathbf{n}}$); that is, perpendicular to the surface.

Since the normal to a flat surface can point in either direction from the surface, the direction of the area vector of an open surface needs to be chosen, as shown in [link].



The direction of the area vector of an open surface needs to be chosen; it could be either of the two cases displayed here. The area vector of a part of a closed surface is defined to point from the inside of the closed space to the outside. This rule gives a unique direction.

Since $\hat{\mathbf{n}}$ is a unit normal to a surface, it has two possible directions at every point on that surface ([link](a)). For an open surface, we can use either direction, as long as we are consistent over the entire surface. Part (c) of the figure shows several cases.



$\hat{n}_1$
$\hat{n}_2$
(a)
$\hat{n}$
$\hat{n}$
$\hat{n}$
$\hat{n}$
$\hat{n}$
(b)
$\hat{n}$
$\hat{n}$
S_1
No
$\hat{n}$
$\hat{n}$
$\hat{n}$
S_2
No
$\hat{n}$
$\hat{n}$
$\hat{n}$
S_3
Yes
(c)

(a) Two potential normal vectors arise at every point on a surface. (b) The outward normal is used to calculate the flux through a closed surface. (c)

Only S_3 has been given a consistent set of normal vectors that allows us to define the flux through the surface.

However, if a surface is closed, then the surface encloses a volume. In that case, the direction of the normal vector at any point on the surface points from the inside to the outside. On a *closed surface* such as that of [link](b), $\mathbf{\hat{n}}$ is chosen to be the *outward normal* at every point, to be consistent with the sign convention for electric charge.

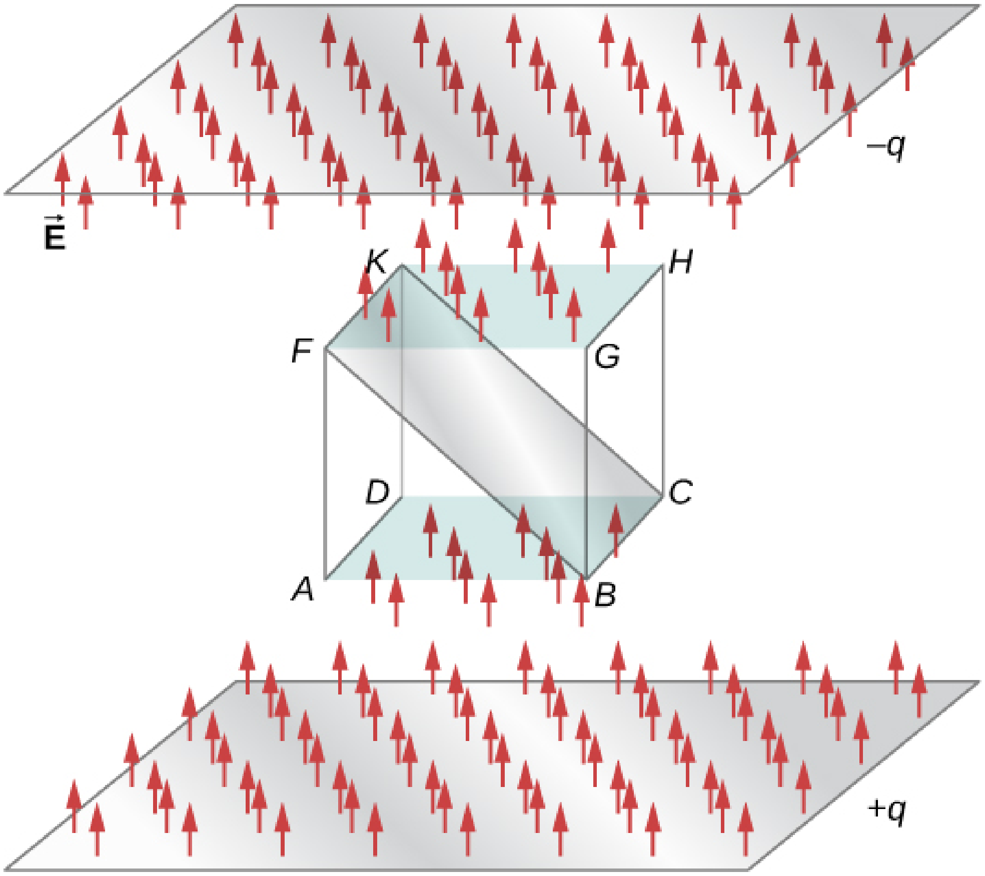
Electric Flux

Now that we have defined the area vector of a surface, we can define the electric flux of a uniform electric field through a flat area as the scalar product of the electric field and the area vector, as defined in Products of Vectors:

Note:
Equation:

$$\Phi = \overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{A}}\,(\text{uniform}\,\overrightarrow{\mathbf{E}},\ \text{flat surface}).$$

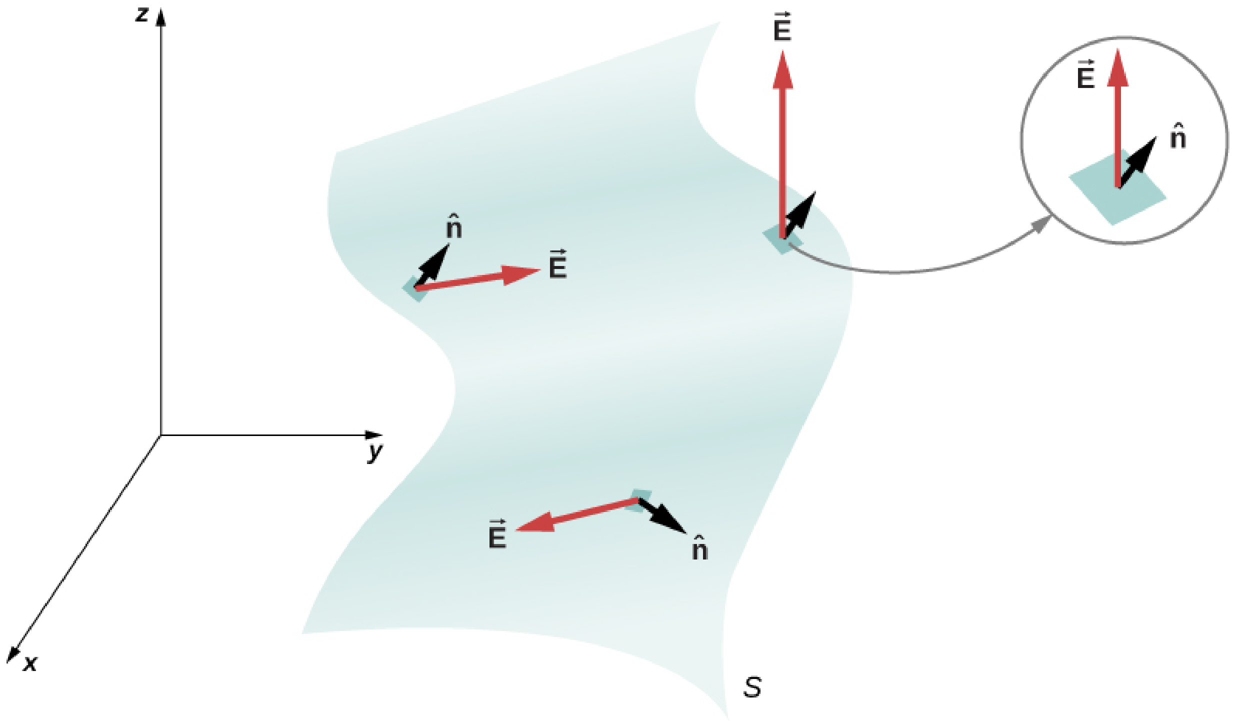
[link] shows the electric field of an oppositely charged, parallel-plate system and an imaginary box between the plates. The electric field between the plates is uniform and points from the positive plate toward the negative plate. A calculation of the flux of this field through various faces of the box shows that the net flux through the box is zero. Why does the flux cancel out here?



Electric flux through a cube, placed between two charged plates. Electric flux through the bottom face (*ABCD*) is negative, because $\vec{\mathbf{E}}$ is in the opposite direction to the normal to the surface. The electric flux through the top face (*FGHK*) is positive, because the electric field and the normal are in the same direction. The electric flux through the other faces is zero, since the electric field is perpendicular to the normal vectors of those faces. The net electric flux through the cube is the sum of fluxes through the six faces. Here, the net flux through the cube is equal to zero. The magnitude of the flux through rectangle *BCKF* is equal to the magnitudes of the flux through both the top and bottom faces.

The reason is that the sources of the electric field are outside the box. Therefore, if any electric field line enters the volume of the box, it must also exit somewhere on the surface because there is no charge inside for the lines to land on. Therefore, quite generally, electric flux through a closed surface is zero if there are no sources of electric field, whether positive or negative charges, inside the enclosed volume. In general, when field lines leave (or “flow out of”) a closed surface, Φ is positive; when they enter (or “flow into”) the surface, Φ is negative.

Any smooth, non-flat surface can be replaced by a collection of tiny, approximately flat surfaces, as shown in [link]. If we divide a surface S into small patches, then we notice that, as the patches become smaller, they can be approximated by flat surfaces. This is similar to the way we treat the surface of Earth as locally flat, even though we know that globally, it is approximately spherical.



A surface is divided into patches to find the flux.

To keep track of the patches, we can number them from 1 through N. Now, we define the area vector for each patch as the area of the patch pointed in the direction of the normal. Let us denote the area vector for the ith patch by $\delta \overrightarrow{\mathbf{A}}_i$. (We have used the symbol δ to remind us that the area is of an arbitrarily small patch.) With sufficiently small patches, we may approximate the electric field over any given patch as uniform. Let us denote the average electric field at the location of the *i*th patch by $\vec{\mathbf{E}}_i$.

Equation:

$$\vec{\mathbf{E}}_i = \text{average electric field over the } i\text{th patch.}$$

Therefore, we can write the electric flux Φ_i through the area of the *i*th patch as

Equation:

$$\Phi_i = \overrightarrow{\mathbf{E}}_i \cdot \delta\overrightarrow{\mathbf{A}}_i \; (i\text{th patch}).$$

The flux through each of the individual patches can be constructed in this manner and then added to give us an estimate of the net flux through the entire surface *S*, which we denote simply as Φ.

Equation:

$$\Phi = \sum_{i=1}^{N} \Phi_i = \sum_{i=1}^{N} \overrightarrow{\mathbf{E}}_i \cdot \delta\overrightarrow{\mathbf{A}}_i \; (N \text{ patch estimate}).$$

This estimate of the flux gets better as we decrease the size of the patches. However, when you use smaller patches, you need more of them to cover the same surface. In the limit of infinitesimally small patches, they may be considered to have area *dA* and unit normal $\hat{\mathbf{n}}$. Since the elements are infinitesimal, they may be assumed to be planar, and $\overrightarrow{\mathbf{E}}_i$ may be taken as constant over any element. Then the flux $d\Phi$ through an area *dA* is given by $d\Phi = \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA$. It is positive when the angle between $\overrightarrow{\mathbf{E}}_i$ and $\hat{\mathbf{n}}$ is less than 90° and negative when the angle is greater than 90°. The net flux is the sum of the infinitesimal flux elements over the entire surface. With infinitesimally small patches, you need infinitely many patches, and the limit of the sum becomes a surface integral. With $\int_S$ representing the integral over *S*,

Note:
Equation:

$$\Phi = \int_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \int_S \overrightarrow{\mathbf{E}} \cdot d\overrightarrow{\mathbf{A}} \text{ (open surface).}$$

In practical terms, surface integrals are computed by taking the antiderivatives of both dimensions defining the area, with the edges of the surface in question being the bounds of the integral.

To distinguish between the flux through an open surface like that of [link] and the flux through a closed surface (one that completely bounds some volume), we represent flux through a closed surface by

Note:
Equation:

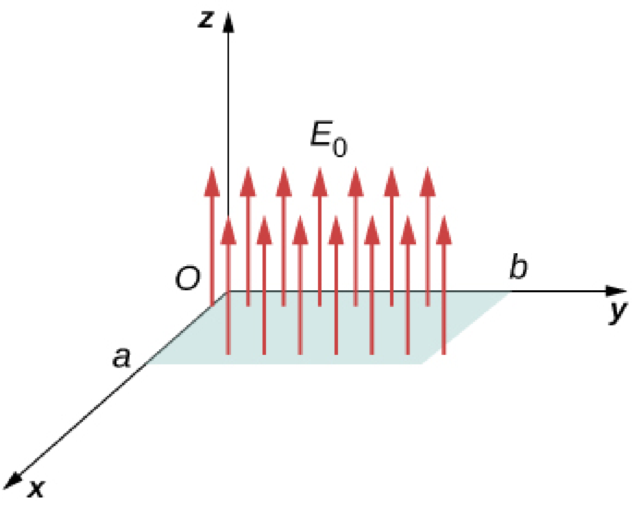
$$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \oint_S \overrightarrow{\mathbf{E}} \cdot d\overrightarrow{\mathbf{A}} \text{ (closed surface)}$$

where the circle through the integral symbol simply means that the surface is closed, and we are integrating over the entire thing. If you only integrate over a portion of a closed surface, that means you are treating a subset of it as an open surface.

Example:

Flux of a Uniform Electric Field

A constant electric field of magnitude E_0 points in the direction of the positive *z*-axis ([link]). What is the electric flux through a rectangle with sides *a* and *b* in the (a) *xy*-plane and in the (b) *xz*-plane?



Calculating the flux of E_0 through a rectangular surface.

Strategy

Apply the definition of flux: $\Phi = \vec{\mathbf{E}} \cdot \vec{\mathbf{A}}$ (uniform $\vec{\mathbf{E}}$), where the definition of dot product is crucial.

Solution

a. In this case, $\Phi = \vec{\mathbf{E}}_0 \cdot \vec{\mathbf{A}} = E_0 A = E_0 ab$.
b. Here, the direction of the area vector is either along the positive *y*-axis or toward the negative *y*-axis. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

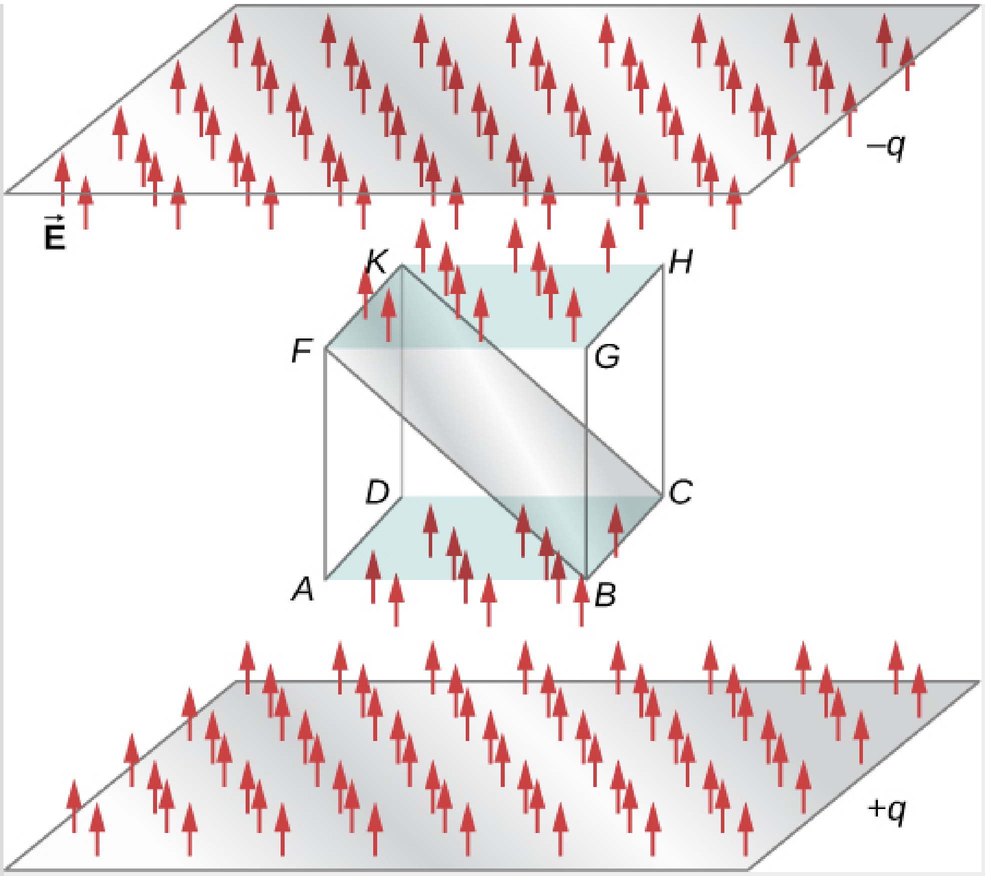
Significance

The relative directions of the electric field and area can cause the flux through the area to be zero.

Example:

Flux of a Uniform Electric Field through a Closed Surface

A constant electric field of magnitude E_0 points in the direction of the positive *z*-axis ([link]). What is the net electric flux through a cube?



Calculating the flux of E_0 through a closed cubic surface.

Strategy

Apply the definition of flux: $\Phi = \overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{A}}$ (uniform $\overrightarrow{\mathbf{E}}$), noting that a closed surface eliminates the ambiguity in the direction of the area vector.

Solution

Through the top face of the cube, $\Phi = \vec{\mathbf{E}}_0 \cdot \vec{\mathbf{A}} = E_0 A$.

Through the bottom face of the cube, $\Phi = \vec{\mathbf{E}}_0 \cdot \vec{\mathbf{A}} = -E_0 A$, because the area vector here points downward.

Along the other four sides, the direction of the area vector is perpendicular to the direction of the electric field. Therefore, the scalar product of the electric field with the area vector is zero, giving zero flux.

The net flux is $\Phi_{\text{net}} = E_0 A - E_0 A + 0 + 0 + 0 + 0 = 0$.

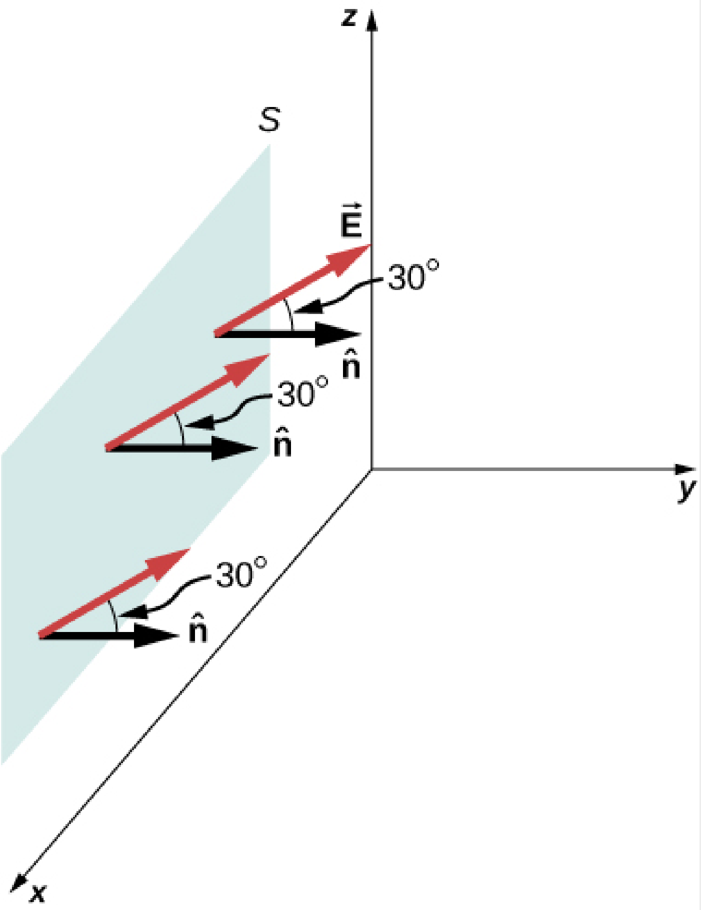
Significance

The net flux of a uniform electric field through a closed surface is zero.

Example:

Electric Flux through a Plane, Integral Method

A uniform electric field $\vec{\mathbf{E}}$ of magnitude 10 N/C is directed parallel to the *yz*-plane at 30° above the *xy*-plane, as shown in [link]. What is the electric flux through the plane surface of area $6.0\,\text{m}^2$ located in the *xz*-plane? Assume that $\hat{\mathbf{n}}$ points in the positive *y*-direction.



The electric field produces a net electric flux through the surface *S*.

Strategy

Apply $\Phi = \int_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA$, where the direction and magnitude of the electric field are constant.

Solution

The angle between the uniform electric field $\vec{\mathbf{E}}$ and the unit normal $\hat{\mathbf{n}}$ to the planar surface is 30°. Since both the direction and magnitude are constant, E comes outside the integral. All that is left is a surface integral over dA, which is A. Therefore, using the open-surface equation, we find that the electric flux through the surface is

Equation:

$$\begin{aligned}\Phi &= \int_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = EA\cos\theta \\ &= (10\,\text{N/C})(6.0\,\text{m}^2)(\cos 30^\circ) = 52\,\text{N}\cdot\text{m}^2/\text{C}.\end{aligned}$$

Significance

Again, the relative directions of the field and the area matter, and the general equation with the integral will simplify to the simple dot product of area and electric field.

Note:

Exercise:

Problem:

Check Your Understanding What angle should there be between the electric field and the surface shown in [link] in the previous example so that no electric flux passes through the surface?

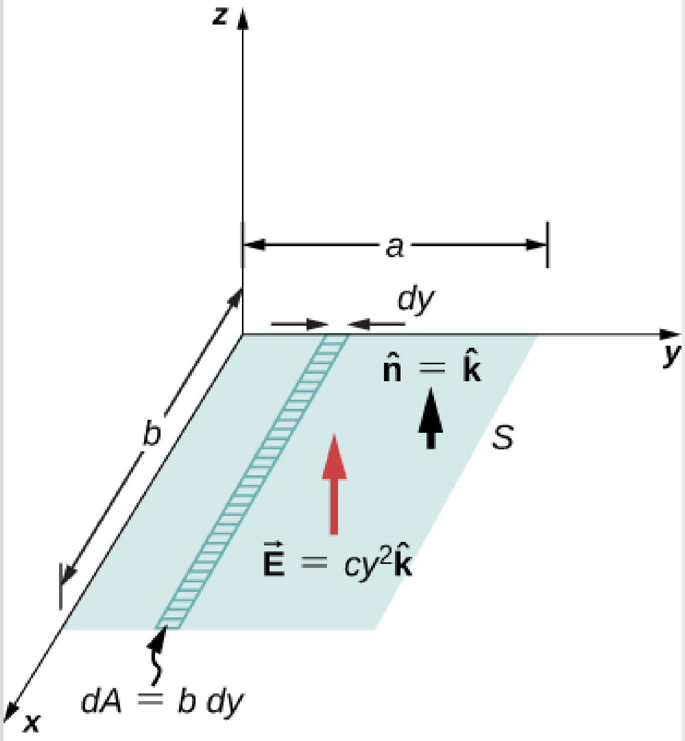
Solution:

Place it so that its unit normal is perpendicular to $\vec{\mathbf{E}}$.

Example:

Inhomogeneous Electric Field

What is the total flux of the electric field $\vec{\mathbf{E}} = cy^2\hat{\mathbf{k}}$ through the rectangular surface shown in [link]?



z
a
dy
y
$\hat{\mathbf{n}} = \hat{\mathbf{k}}$
b
S
$\vec{\mathbf{E}} = cy^2\hat{\mathbf{k}}$
$dA = b\,dy$
x

Since the electric field is not constant over the surface, an integration is necessary to determine the flux.

Strategy

Apply $\Phi = \int_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA$. We assume that the unit normal $\hat{\mathbf{n}}$ to the given surface points in the positive *z*-direction, so $\hat{\mathbf{n}} = \hat{\mathbf{k}}$. Since the electric field is not uniform over the surface, it is necessary to divide the surface into infinitesimal strips along which $\vec{\mathbf{E}}$ is essentially constant. As shown in [link], these strips are parallel to the *x*-axis, and each strip has an area $dA = b\,dy$.

Solution

From the open surface integral, we find that the net flux through the rectangular surface is

Equation:

$$\begin{aligned}\Phi &= \int_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \int_0^a (cy^2 \hat{\mathbf{k}}) \cdot \hat{\mathbf{k}}(b\, dy) \\ &= cb \int_0^a y^2\, dy = \frac{1}{3}\, a^3 bc.\end{aligned}$$

Significance
For a non-constant electric field, the integral method is required.

Note:
Exercise:

Problem:

Check Your Understanding If the electric field in [link] is $\vec{\mathbf{E}} = mx\hat{\mathbf{k}}$, what is the flux through the rectangular area?

Solution:

$mab^2/2$

Summary

- The electric flux through a surface is proportional to the number of field lines crossing that surface. Note that this means the magnitude is proportional to the portion of the field perpendicular to the area.
- The electric flux is obtained by evaluating the surface integral
  Equation:

  $$\Phi = \oint_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \oint_S \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}},$$

  where the notation used here is for a closed surface *S*.

Conceptual Questions

Exercise:

Problem:

Discuss how to orient a planar surface of area A in a uniform electric field of magnitude E_0 to obtain (a) the maximum flux and (b) the minimum flux through the area.

Solution:

a. If the planar surface is perpendicular to the electric field vector, the maximum flux would be obtained. b. If the planar surface were parallel to the electric field vector, the minimum flux would be obtained.

Exercise:

Problem:

What are the maximum and minimum values of the flux in the preceding question?

Exercise:

Problem:

The net electric flux crossing a closed surface is always zero. True or false?

Solution:

False. The net electric flux crossing a closed surface is always zero if and only if the net charge enclosed is zero.

Exercise:

Problem:

The net electric flux crossing an open surface is never zero. True or false?

Problems

Exercise:

Problem:

A uniform electric field of magnitude $1.1 \times 10^4\ \mathrm{N/C}$ is perpendicular to a square sheet with sides 2.0 m long. What is the electric flux through the sheet?

Exercise:

Problem:

Calculate the flux through the sheet of the previous problem if the plane of the sheet is at an angle of 60° to the field. Find the flux for both directions of the unit normal to the sheet.

Solution:

$\Phi = \overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{A}} \to EA\cos\theta = 2.2 \times 10^4 \mathrm{N} \cdot \mathrm{m}^2/\mathrm{C}$ electric field in direction of unit normal; $\Phi = \overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{A}} \to EA\cos\theta = -2.2 \times 10^4 \mathrm{N} \cdot \mathrm{m}^2/\mathrm{C}$ electric field opposite to unit normal

Exercise:

Problem:

Find the electric flux through a rectangular area $3\ \mathrm{cm} \times 2\ \mathrm{cm}$ between two parallel plates where there is a constant electric field of 30 N/C for the following orientations of the area: (a) parallel to the plates, (b) perpendicular to the plates, and (c) the normal to the area making a 30° angle with the direction of the electric field. Note that this angle can also be given as $180^\circ + 30^\circ$.

Exercise:

Problem:

The electric flux through a square-shaped area of side 5 cm near a large charged sheet is found to be $3 \times 10^{-5}\ \mathrm{N} \cdot \mathrm{m}^2/\mathrm{C}$ when the area is parallel to the plate. Find the charge density on the sheet.

Solution:

$$\frac{3 \times 10^{-5}\ \mathrm{N{\cdot}m^2/C}}{(0.05\ \mathrm{m})^2} = E \Rightarrow \sigma = 2.12 \times 10^{-13}\ \mathrm{C/m^2}$$

Exercise:

Problem:

Two large rectangular aluminum plates of area $150\,\text{cm}^2$ face each other with a separation of 3 mm between them. The plates are charged with equal amount of opposite charges, $\pm 20\,\mu\text{C}$. The charges on the plates face each other. Find the flux through a circle of radius 3 cm between the plates when the normal to the circle makes an angle of 5° with a line perpendicular to the plates. Note that this angle can also be given as $180^\circ + 5^\circ$.

Exercise:

Problem:

A square surface of area $2\,\text{cm}^2$ is in a space of uniform electric field of magnitude $10^3\,\text{N/C}$. The amount of flux through it depends on how the square is oriented relative to the direction of the electric field. Find the electric flux through the square, when the normal to it makes the following angles with electric field: (a) 30°, (b) 90°, and (c) 0°. Note that these angles can also be given as $180^\circ + \theta$.

Solution:

a. $\Phi = 0.17\,\text{N} \cdot \text{m}^2/\text{C}$;
b. $\Phi = 0$; c.
$\Phi = EA \cos 0^\circ = 1.0 \times 10^3\,\text{N/C}(2.0 \times 10^{-4}\,\text{m})^2 \cos 0^\circ = 0.20\,\text{N} \cdot \text{m}^2/\text{C}$

Exercise:

Problem:

A vector field is pointed along the *z*-axis, $\vec{\mathbf{v}} = \frac{\alpha}{x^2+y^2}\,\hat{\mathbf{z}}$. (a) Find the flux of the vector field through a rectangle in the *xy*-plane between $a < x < b$ and $c < y < d$. (b) Do the same through a rectangle in the *yz*-plane between $a < z < b$ and $c < y < d$. (Leave your answer as an integral.)

Exercise:

Problem:

Consider the uniform electric field $\vec{\mathbf{E}} = (4.0\,\hat{\mathbf{j}} + 3.0\,\hat{\mathbf{k}}) \times 10^3\,\text{N/C}$. What is its electric flux through a circular area of radius 2.0 m that lies in the *xy*-plane?

Solution:

$$\Phi = 3.8 \times 10^4 \, \mathrm{N \cdot m^2/C}$$

Exercise:

Problem:

Repeat the previous problem, given that the circular area is (a) in the *yz*-plane and (b) 45° above the *xy*-plane.

Exercise:

Problem:

An infinite charged wire with charge per unit length λ lies along the central axis of a cylindrical surface of radius *r* and length *l*. What is the flux through the surface due to the electric field of the charged wire?

Solution:

$$\vec{\mathbf{E}}(z) = \frac{1}{4\pi\varepsilon_0} \frac{2\lambda}{z} \hat{\mathbf{k}}, \; \int \vec{\mathbf{E}} \cdot \hat{\mathbf{n}} \, dA = \frac{\lambda}{\varepsilon_0} l$$

Glossary

area vector
vector with magnitude equal to the area of a surface and direction perpendicular to the surface

electric flux
dot product of the electric field and the area through which it is passing

flux
quantity of something passing through a given area

Explaining Gauss's Law

By the end of this section, you will be able to:

- State Gauss's law
- Explain the conditions under which Gauss's law may be used
- Apply Gauss's law in appropriate systems

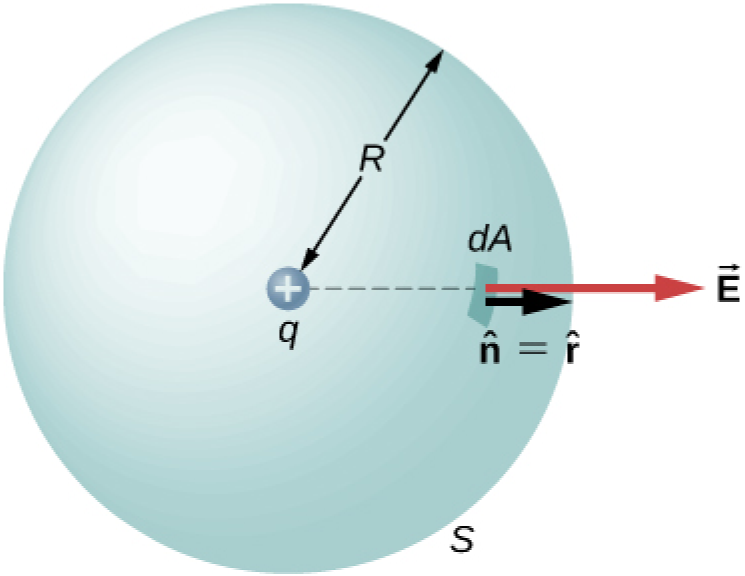
We can now determine the electric flux through an arbitrary closed surface due to an arbitrary charge distribution. We found that if a closed surface does not have any charge inside where an electric field line can terminate, then any electric field line entering the surface at one point must necessarily exit at some other point of the surface. Therefore, if a closed surface does not have any charges inside the enclosed volume, then the electric flux through the surface is zero. Now, what happens to the electric flux if there are some charges inside the enclosed volume? Gauss's law gives a quantitative answer to this question.

To get a feel for what to expect, let's calculate the electric flux through a spherical surface around a positive point charge *q*, since we already know the electric field in such a situation. Recall that when we place the point charge at the origin of a coordinate system, the electric field at a point *P* that is at a distance *r* from the charge at the origin is given by

Equation:

$$\vec{\mathbf{E}}_P = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}},$$

where $\hat{\mathbf{r}}$ is the radial vector from the charge at the origin to the point *P*. We can use this electric field to find the flux through the spherical surface of radius *r*, as shown in [link].



A closed spherical surface surrounding a point charge *q*.

Then we apply $\Phi = \int_S \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA$ to this system and substitute known values. On the sphere, $\widehat{\mathbf{n}} = \hat{\mathbf{r}}$ and $r = R$, so for an infinitesimal area *dA*,

Equation:

$$d\Phi = \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = \frac{1}{4\pi\varepsilon_0}\ \frac{q}{R^2} \hat{\mathbf{r}} \cdot \hat{\mathbf{r}}\, dA = \frac{1}{4\pi\varepsilon_0}\ \frac{q}{R^2}\, dA.$$

We now find the net flux by integrating this flux over the surface of the sphere:

Equation:

$$\Phi = \frac{1}{4\pi\varepsilon_0}\ \frac{q}{R^2} \oint_S dA = \frac{1}{4\pi\varepsilon_0}\ \frac{q}{R^2} (4\pi R^2) = \frac{q}{\varepsilon_0}.$$

where the total surface area of the spherical surface is $4\pi R^2$. This gives the flux through the closed spherical surface at radius *r* as

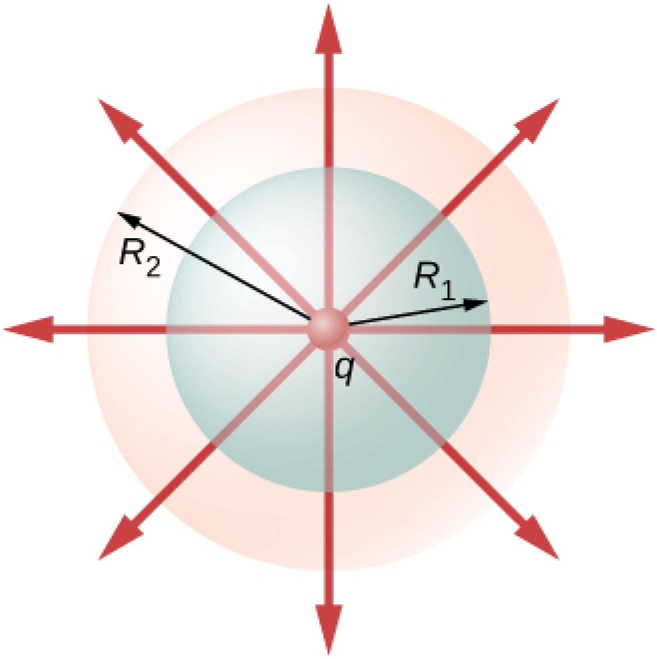
Equation:

$$\Phi = \frac{q}{\varepsilon_0}.$$

A remarkable fact about this equation is that the flux is independent of the size of the spherical surface. This can be directly attributed to the fact that the electric field of a point charge decreases as $1/r^2$ with distance, which just cancels the r^2 rate of increase of the surface area.

Electric Field Lines Picture

An alternative way to see why the flux through a closed spherical surface is independent of the radius of the surface is to look at the electric field lines. Note that every field line from q that pierces the surface at radius R_1 also pierces the surface at R_2 ([link]).

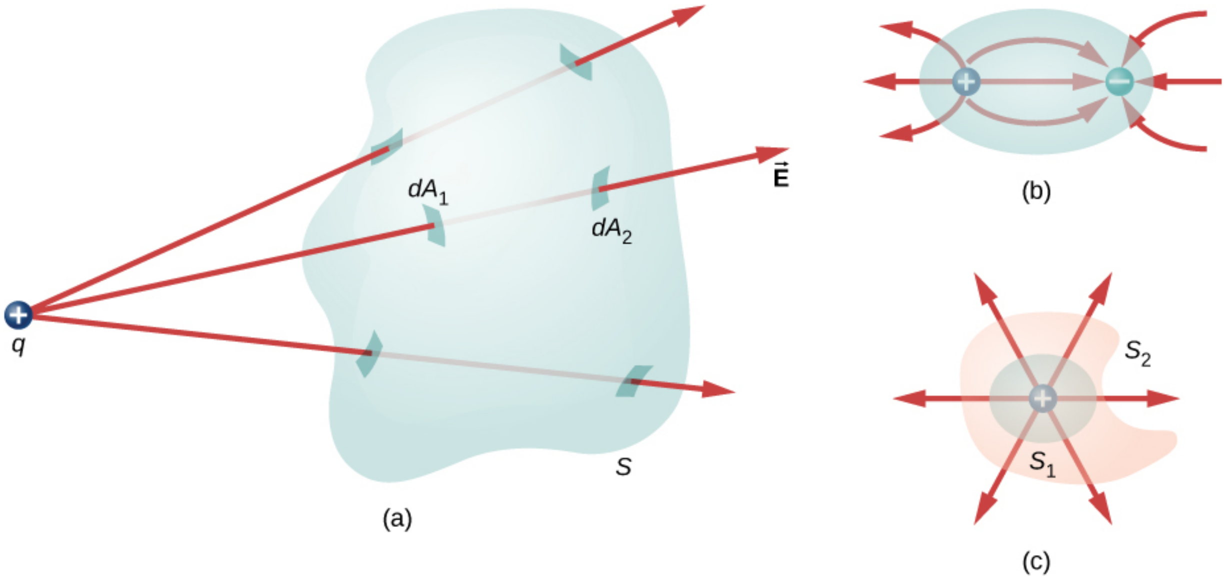


Flux through spherical surfaces of radii R_1 and R_2 enclosing a charge

q are equal, independent of the size of the surface, since all E-field lines that pierce one surface from the inside to outside direction also pierce the other surface in the same direction.

Therefore, the net number of electric field lines passing through the two surfaces from the inside to outside direction is equal. This net number of electric field lines, which is obtained by subtracting the number of lines in the direction from outside to inside from the number of lines in the direction from inside to outside gives a visual measure of the electric flux through the surfaces.

You can see that if no charges are included within a closed surface, then the electric flux through it must be zero. A typical field line enters the surface at dA_1 and leaves at dA_2. Every line that enters the surface must also leave that surface. Hence the net “flow” of the field lines into or out of the surface is zero ([link](a)). The same thing happens if charges of equal and opposite sign are included inside the closed surface, so that the total charge included is zero (part (b)). A surface that includes the same amount of charge has the same number of field lines crossing it, regardless of the shape or size of the surface, as long as the surface encloses the same amount of charge (part (c)).



Understanding the flux in terms of field lines. (a) The electric flux through a closed surface due to a charge outside that surface is zero. (b) Charges are enclosed, but because the net charge included is zero, the net flux through the closed surface is also zero. (c) The shape and size of the surfaces that enclose a charge does not matter because all surfaces enclosing the same charge have the same flux.

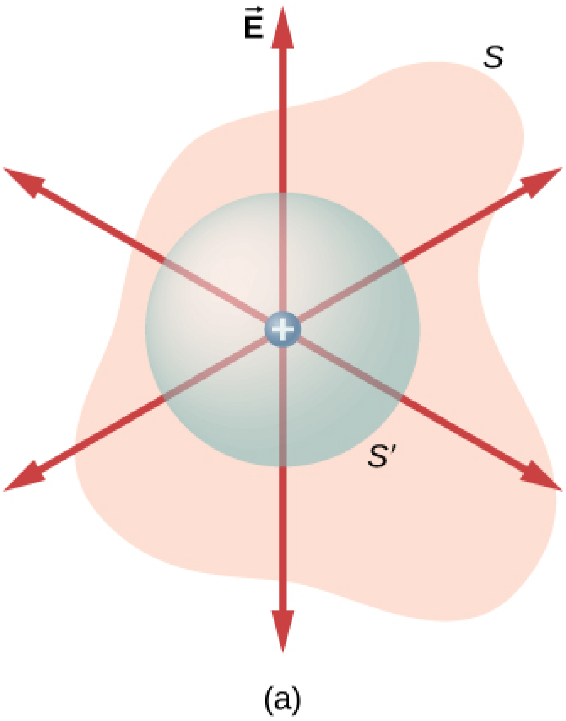
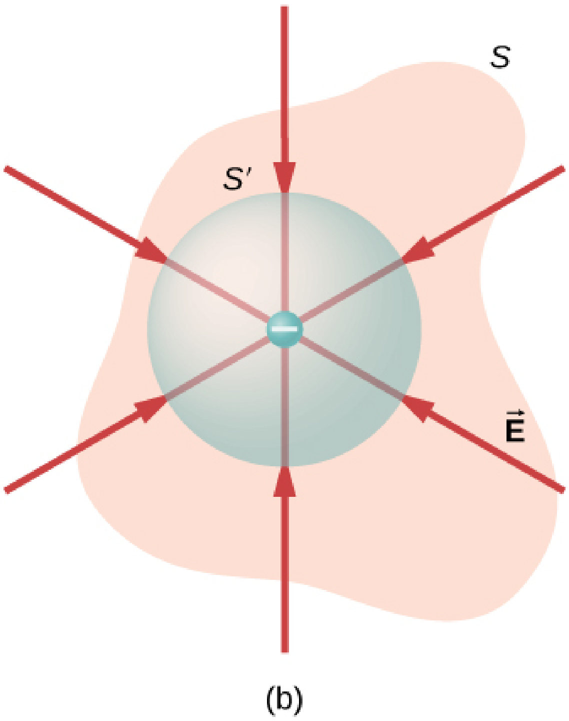
Statement of Gauss’s Law

Gauss’s law generalizes this result to the case of any number of charges and any location of the charges in the space inside the closed surface. According to Gauss’s law, the flux of the electric field $\overrightarrow{\mathbf{E}}$ through any closed surface, also called a **Gaussian surface**, is equal to the net charge enclosed (q_{enc}) divided by the permittivity of free space (ε_0):

Equation:

$$\Phi_{\text{Closed Surface}} = \frac{q_{\text{enc}}}{\varepsilon_0}.$$

This equation holds for *charges of either sign*, because we define the area vector of a closed surface to point outward. If the enclosed charge is negative (see [link](b)), then the flux through either S or S' is negative.

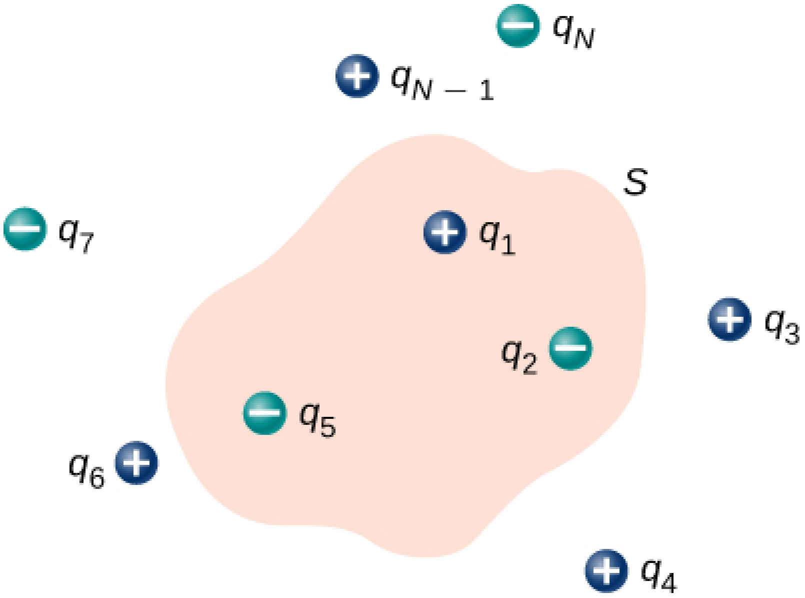



The electric flux through any closed surface surrounding a point charge *q* is given by Gauss's law. (a) Enclosed charge is positive. (b) Enclosed charge is negative.

The Gaussian surface does not need to correspond to a real, physical object; indeed, it rarely will. It is a mathematical construct that may be of any shape, provided that it is closed. However, since our goal is to integrate the flux over it, we tend to choose shapes that are highly symmetrical.

If the charges are discrete point charges, then we just add them. If the charge is described by a continuous distribution, then we need to integrate appropriately to find the total charge that resides inside the enclosed volume. For example, the flux through the Gaussian surface *S* of [link] is

$\Phi = (q_1 + q_2 + q_5)/\varepsilon_0$. Note that q_{enc} is simply the sum of the point charges. If the charge distribution were continuous, we would need to integrate appropriately to compute the total charge within the Gaussian surface.



The flux through the Gaussian surface shown, due to the charge distribution, is $\Phi = |q_1| + |q_2| + |q_5|/\varepsilon_0$.

Recall that the principle of superposition holds for the electric field. Therefore, the total electric field at any point, including those on the chosen Gaussian surface, is the sum of all the electric fields present at this point. This allows us to write Gauss's law in terms of the total electric field.

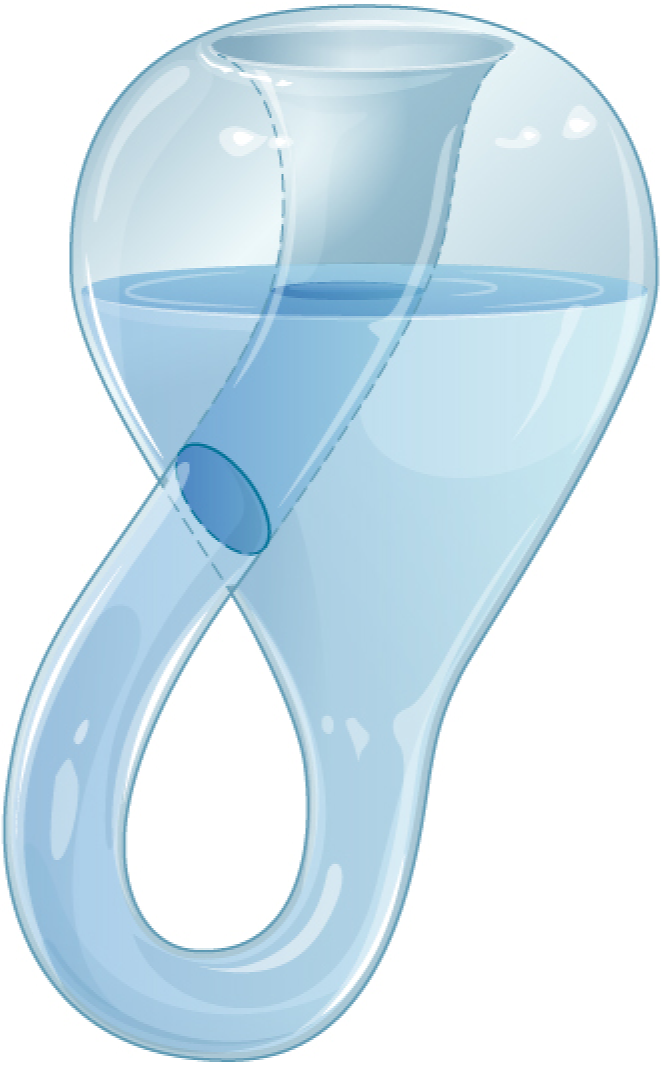
Note:
Gauss's Law

The flux Φ of the electric field $\overrightarrow{\mathbf{E}}$ through any closed surface S (a Gaussian surface) is equal to the net charge enclosed $\left(q_{\text{enc}}\right)$ divided by the permittivity of free space $\left(\varepsilon_0\right)$:

Equation:

$$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = \frac{q_{\text{enc}}}{\varepsilon_0}.$$

To use Gauss's law effectively, you must have a clear understanding of what each term in the equation represents. The field $\overrightarrow{\mathbf{E}}$ is the *total electric field* at every point on the Gaussian surface. This total field includes contributions from charges both inside and outside the Gaussian surface. However, q_{enc} is just the charge *inside* the Gaussian surface. Finally, the Gaussian surface is any closed surface in space. That surface can coincide with the actual surface of a conductor, or it can be an imaginary geometric surface. The only requirement imposed on a Gaussian surface is that it be closed ([link]).

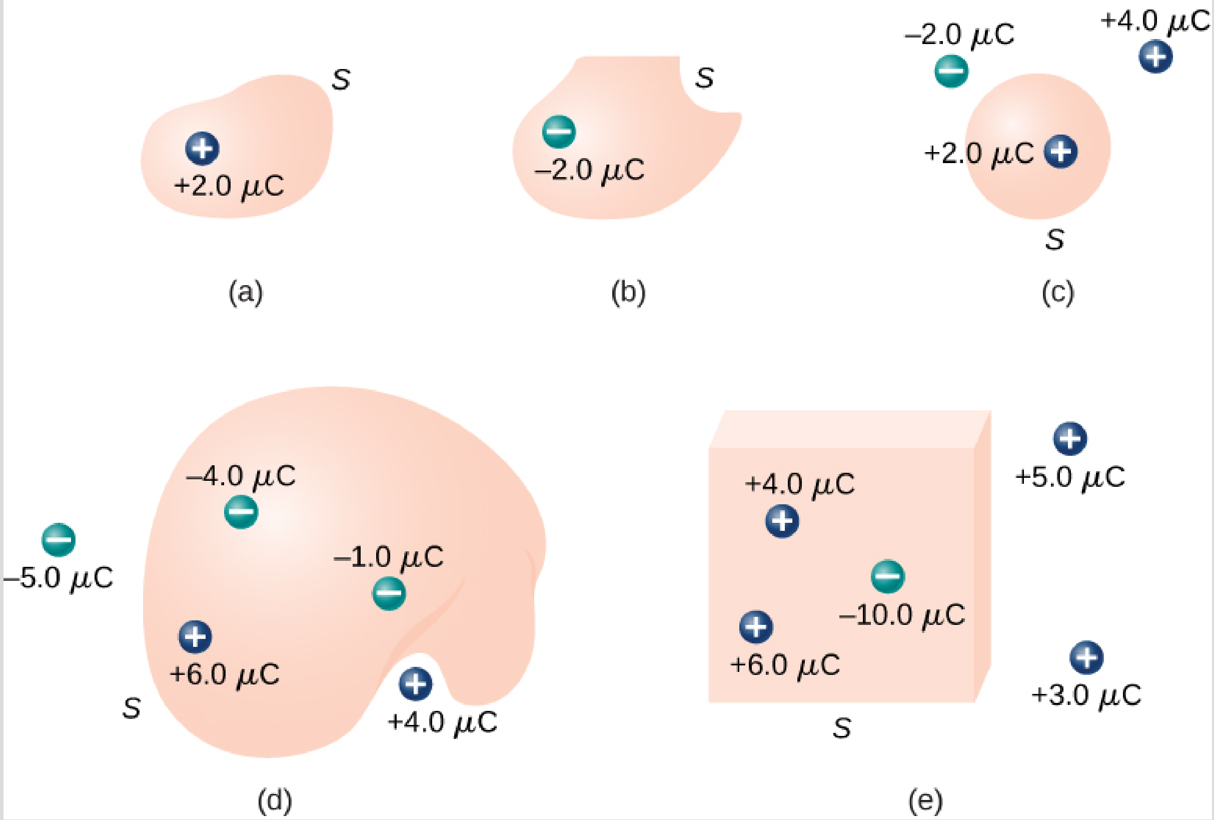


A Klein bottle partially filled with a liquid. Could the Klein bottle be used as a Gaussian surface?

Example:

Electric Flux through Gaussian Surfaces

Calculate the electric flux through each Gaussian surface shown in [link].


S
+2.0 μC
(a)
S
–2.0 μC
(b)
–2.0 μC
+4.0 μC
+2.0 μC
S
(c)
–4.0 μC
–5.0 μC
–1.0 μC
+6.0 μC
S
+4.0 μC
(d)
+4.0 μC
+5.0 μC
–10.0 μC
+6.0 μC
+3.0 μC
S
(e)

Various Gaussian surfaces and charges.

Strategy

From Gauss's law, the flux through each surface is given by $q_{\text{enc}}/\varepsilon_0$, where q_{enc} is the charge enclosed by that surface.

Solution

For the surfaces and charges shown, we find

a. $\Phi = \frac{2.0\,\mu\text{C}}{\varepsilon_0} = 2.3 \times 10^5\,\text{N} \cdot \text{m}^2/\text{C}.$

b. $\Phi = \frac{-2.0\,\mu\text{C}}{\varepsilon_0} = -2.3 \times 10^5\,\text{N} \cdot \text{m}^2/\text{C}.$

c. $\Phi = \frac{2.0\,\mu\text{C}}{\varepsilon_0} = 2.3 \times 10^5\,\text{N} \cdot \text{m}^2/\text{C}.$

d. $\Phi = \frac{-4.0\,\mu\text{C}+6.0\,\mu\text{C}-1.0\,\mu\text{C}}{\varepsilon_0} = 1.1 \times 10^5\,\text{N} \cdot \text{m}^2/\text{C}.$

e. $\Phi = \frac{4.0\,\mu\text{C}+6.0\,\mu\text{C}-10.0\,\mu\text{C}}{\varepsilon_0} = 0.$

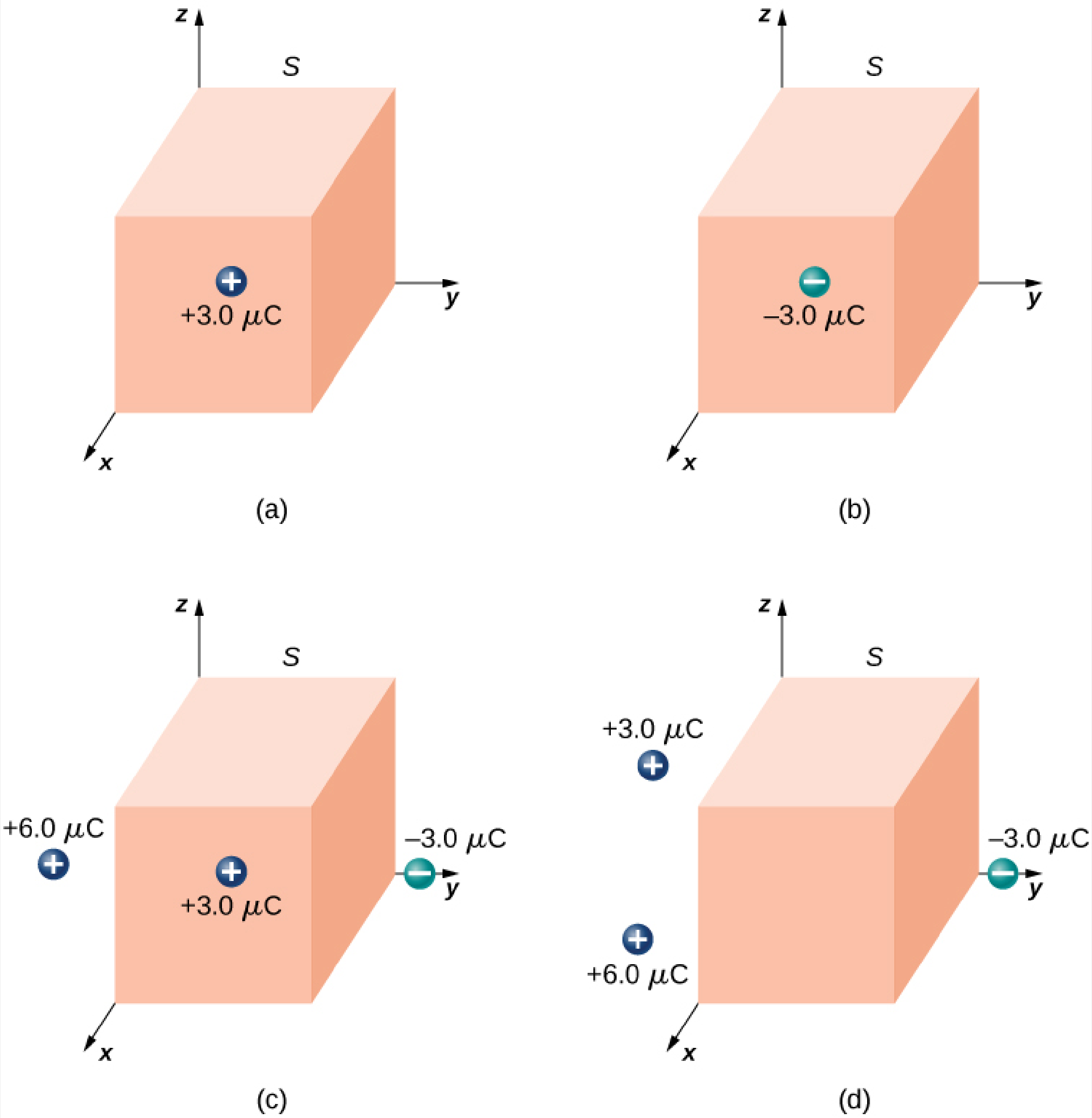
Significance

In the special case of a closed surface, the flux calculations become a sum of charges. In the next section, this will allow us to work with more complex systems.

Note:
Exercise:

Problem:

Check Your Understanding Calculate the electric flux through the closed cubical surface for each charge distribution shown in [link].


z
S
y
x
+3.0 μC
(a)
z
S
y
x
–3.0 μC
(b)
z
S
+6.0 μC
+3.0 μC
–3.0 μC
y
x
(c)
z
S
+3.0 μC
–3.0 μC
y
+6.0 μC
x
(d)

A cubical Gaussian surface with various charge distributions.

Solution:

a. $3.4 \times 10^5 \, \text{N} \cdot \text{m}^2/\text{C}$; b. $-3.4 \times 10^5 \, \text{N} \cdot \text{m}^2/\text{C}$; c. $3.4 \times 10^5 \, \text{N} \cdot \text{m}^2/\text{C}$; d. 0

Note:
Use this simulation to adjust the magnitude of the charge and the radius of the Gaussian surface around it. See how this affects the total flux and the magnitude of the electric field at the Gaussian surface.

Summary

- Gauss's law relates the electric flux through a closed surface to the net charge within that surface,
  Equation:
  $$\Phi = \oint_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}} \, dA = \frac{q_{\text{enc}}}{\varepsilon_0},$$
  where q_{enc} is the total charge inside the Gaussian surface S.
- All surfaces that include the same amount of charge have the same number of field lines crossing it, regardless of the shape or size of the surface, as long as the surfaces enclose the same amount of charge.

Conceptual Questions

Exercise:

Problem:

Two concentric spherical surfaces enclose a point charge q. The radius of the outer sphere is twice that of the inner one. Compare the electric fluxes crossing the two surfaces.

Solution:

Since the electric field vector has a $\frac{1}{r^2}$ dependence, the fluxes are the same since $A = 4\pi r^2$.

Exercise:

Problem:

Compare the electric flux through the surface of a cube of side length a that has a charge q at its center to the flux through a spherical surface of radius a with a charge q at its center.

Exercise:

Problem:

(a) If the electric flux through a closed surface is zero, is the electric field necessarily zero at all points on the surface? (b) What is the net charge inside the surface?

Solution:

a. no; b. zero

Exercise:

Problem:

Discuss how Gauss's law would be affected if the electric field of a point charge did not vary as $1/r^2$.

Exercise:

Problem:

Discuss the similarities and differences between the gravitational field of a point mass *m* and the electric field of a point charge *q*.

Solution:

Both fields vary as $\frac{1}{r^2}$. Because the gravitational constant is so much smaller than $\frac{1}{4\pi\varepsilon_0}$, the gravitational field is orders of magnitude weaker than the electric field. Also, the gravitational flux through a closed surface is zero or positive; however, the electric flux is positive, negative, or zero, depending on the definition of flux for the given situation.

Exercise:

Problem:

Discuss whether Gauss's law can be applied to other forces, and if so, which ones.

Exercise:

Problem:

Is the term $\overrightarrow{\mathbf{E}}$ in Gauss's law the electric field produced by just the charge inside the Gaussian surface?

Solution:

No, it is produced by all charges both inside and outside the Gaussian surface.

Exercise:

Problem:

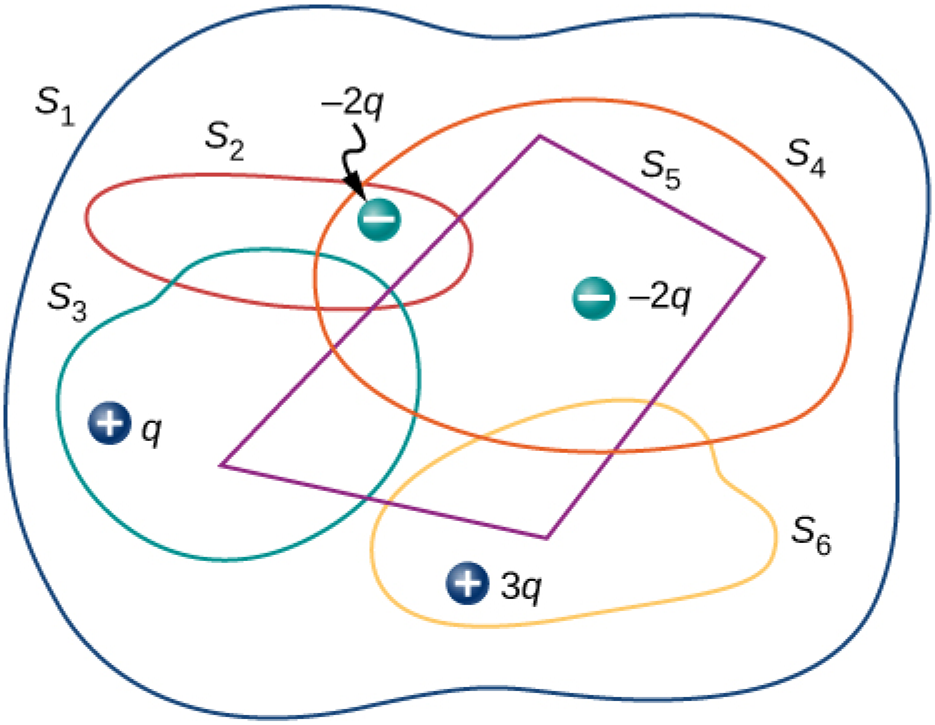
Reformulate Gauss's law by choosing the unit normal of the Gaussian surface to be the one directed inward.

Problems

Exercise:

Problem:

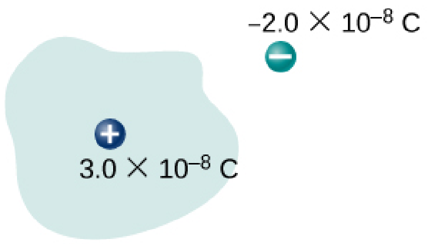
Determine the electric flux through each closed surface where the cross-section inside the surface is shown below.



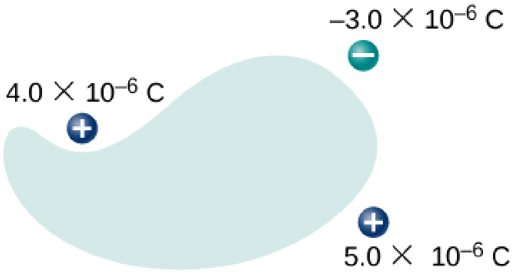
Exercise:

Problem:

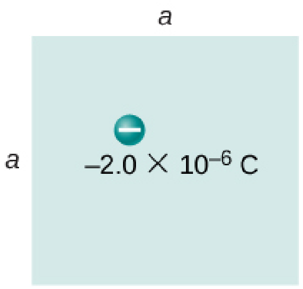
Find the electric flux through the closed surface whose cross-sections are shown below.


–2.0 × 10^{-8} C
3.0 × 10^{-8} C

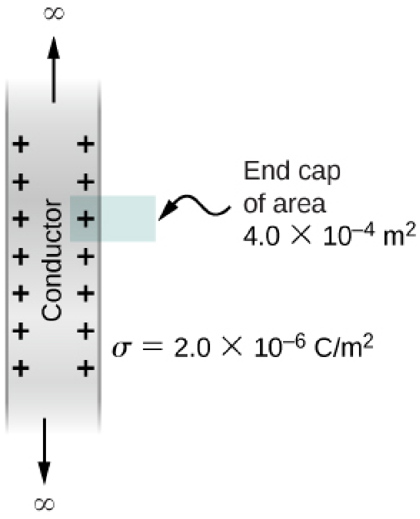
(a)


–3.0 × 10^{-6} C
4.0 × 10^{-6} C
5.0 × 10^{-6} C

(b)


a
a
–2.0 × 10^{-6} C

(c)


∞
Conductor
End cap
of area
4.0 × 10^{-4} m^2
$\sigma = 2.0 \times 10^{-6}$ C/m^2
∞

(d)

Solution:

a. $\Phi = 3.39 \times 10^3 \mathrm{N \cdot m^2/C}$; b. $\Phi = 0$;
c. $\Phi = -2.25 \times 10^5 \mathrm{N \cdot m^2/C}$;
d. $\Phi = 90.4 \mathrm{N \cdot m^2/C}$

Exercise:

Problem:

A point charge q is located at the center of a cube whose sides are of length a. If there are no other charges in this system, what is the electric flux through one face of the cube?

Exercise:

Problem:

A point charge of 10 μC is at an unspecified location inside a cube of side 2 cm. Find the net electric flux though the surfaces of the cube.

Solution:

$\Phi = 1.13 \times 10^6 \mathrm{N \cdot m^2/C}$

Exercise:

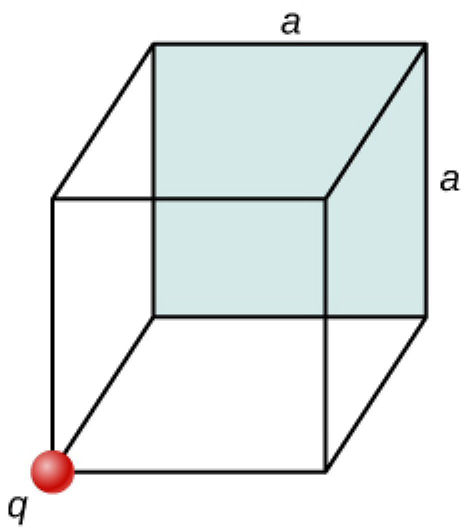
Problem:

A net flux of $1.0 \times 10^4 \mathrm{N \cdot m^2/C}$ passes inward through the surface of a sphere of radius 5 cm. (a) How much charge is inside the sphere? (b) How precisely can we determine the location of the charge from this information?

Exercise:

Problem:

A charge q is placed at one of the corners of a cube of side a, as shown below. Find the magnitude of the electric flux through the shaded face due to q. Assume $q > 0$.



Solution:

Make a cube with *q* at the center, using the cube of side *a*. This would take four cubes of side *a* to make one side of the large cube. The shaded side of the small cube would be 1/24th of the total area of the large cube; therefore, the flux through the shaded area would be $\Phi = \frac{1}{24}\frac{q}{\varepsilon_0}$.

Exercise:

Problem:

The electric flux through a cubical box 8.0 cm on a side is $1.2 \times 10^3\,\text{N}\cdot\text{m}^2/\text{C}$. What is the total charge enclosed by the box?

Exercise:

Problem:

The electric flux through a spherical surface is $4.0 \times 10^4\,\text{N}\cdot\text{m}^2/\text{C}$. What is the net charge enclosed by the surface?

Solution:

$q = 3.54 \times 10^{-7}\,\text{C}$

Exercise:

Problem:

A cube whose sides are of length *d* is placed in a uniform electric field of magnitude $E = 4.0 \times 10^3\ \text{N/C}$ so that the field is perpendicular to two opposite faces of the cube. What is the net flux through the cube?

Exercise:

Problem:

Repeat the previous problem, assuming that the electric field is directed along a body diagonal of the cube.

Solution:

zero, also because flux in equals flux out

Exercise:

Problem:

A total charge $5.0 \times 10^{-6}\ \text{C}$ is distributed uniformly throughout a cubical volume whose edges are 8.0 cm long. (a) What is the charge density in the cube? (b) What is the electric flux through a cube with 12.0-cm edges that is concentric with the charge distribution? (c) Do the same calculation for cubes whose edges are 10.0 cm long and 5.0 cm long. (d) What is the electric flux through a spherical surface of radius 3.0 cm that is also concentric with the charge distribution?

Glossary

Gaussian surface
: any enclosed (usually imaginary) surface

Applying Gauss’s Law

By the end of this section, you will be able to:

- Explain what spherical, cylindrical, and planar symmetry are
- Recognize whether or not a given system possesses one of these symmetries
- Apply Gauss’s law to determine the electric field of a system with one of these symmetries

Gauss’s law is very helpful in determining expressions for the electric field, even though the law is not directly about the electric field; it is about the electric flux. It turns out that in situations that have certain symmetries (spherical, cylindrical, or planar) in the charge distribution, we can deduce the electric field based on knowledge of the electric flux. In these systems, we can find a Gaussian surface *S* over which the electric field has constant magnitude. Furthermore, if $\overrightarrow{\mathbf{E}}$ is parallel to $\widehat{\mathbf{n}}$ everywhere on the surface, then $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}} = E$. (If $\overrightarrow{\mathbf{E}}$ and $\widehat{\mathbf{n}}$ are antiparallel everywhere on the surface, then $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}} = -E$.) Gauss’s law then simplifies to

Note:
Equation:

$$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = E \oint_S dA = EA = \frac{q_{\text{enc}}}{\varepsilon_0},$$

where *A* is the area of the surface. Note that these symmetries lead to the transformation of the flux integral into a product of the magnitude of the electric field and an appropriate area. When you use this flux in the expression for Gauss’s law, you obtain an algebraic equation that you can solve for the magnitude of the electric field, which looks like

Equation:

$$E \sim \frac{q_{\text{enc}}}{\varepsilon_0 \text{ area}}.$$

The direction of the electric field at point *P* is obtained from the symmetry of the charge distribution and the type of charge in the distribution. Therefore, Gauss’s law can be used to determine $\overrightarrow{\mathbf{E}}$. Here is a summary of the steps we will follow:

Note:
Problem-Solving Strategy: Gauss’s Law

1. *Identify the spatial symmetry of the charge distribution*. This is an important first step that allows us to choose the appropriate Gaussian surface. As examples, an isolated point charge

has spherical symmetry, and an infinite line of charge has cylindrical symmetry.

2. *Choose a Gaussian surface with the same symmetry as the charge distribution and identify its consequences*. With this choice, $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}$ is easily determined over the Gaussian surface.
3. *Evaluate the integral* $\oint_S \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA$ *over the Gaussian surface, that is, calculate the flux through the surface*. The symmetry of the Gaussian surface allows us to factor $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}$ outside the integral.
4. *Determine the amount of charge enclosed by the Gaussian surface*. This is an evaluation of the right-hand side of the equation representing Gauss's law. It is often necessary to perform an integration to obtain the net enclosed charge.
5. *Evaluate the electric field of the charge distribution*. The field may now be found using the results of steps 3 and 4.

Basically, there are only three types of symmetry that allow Gauss's law to be used to deduce the electric field. They are

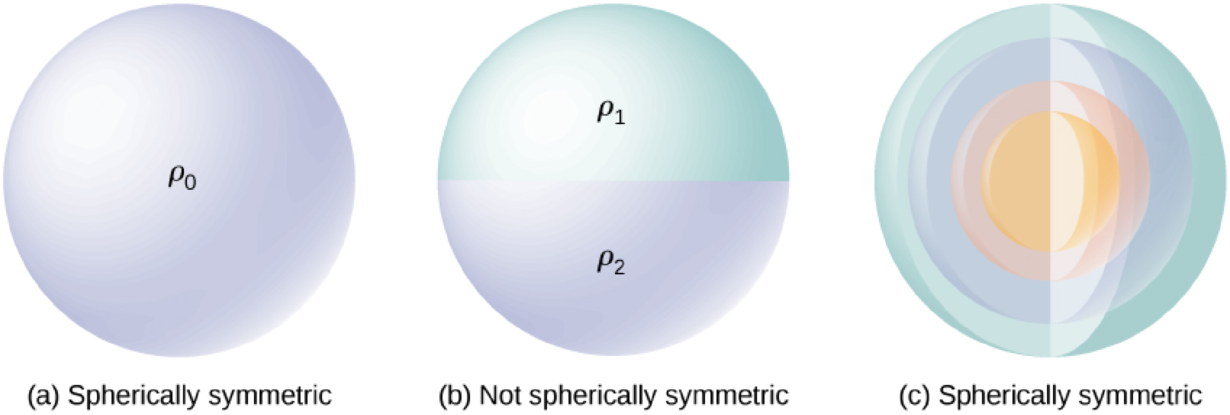
- A charge distribution with spherical symmetry
- A charge distribution with cylindrical symmetry
- A charge distribution with planar symmetry

To exploit the symmetry, we perform the calculations in appropriate coordinate systems and use the right kind of Gaussian surface for that symmetry, applying the remaining four steps.

Charge Distribution with Spherical Symmetry

A charge distribution has **spherical symmetry** if the density of charge depends only on the distance from a point in space and not on the direction. In other words, if you rotate the system, it doesn't look different. For instance, if a sphere of radius R is uniformly charged with charge density ρ_0 then the distribution has spherical symmetry ([link](a)). On the other hand, if a sphere of radius R is charged so that the top half of the sphere has uniform charge density ρ_1 and the bottom half has a uniform charge density $\rho_2 \neq \rho_1$, then the sphere does not have spherical symmetry because the charge density depends on the direction ([link](b)). Thus, it is not the shape of the object but rather the shape of the charge distribution that determines whether or not a system has spherical symmetry.

[link](c) shows a sphere with four different shells, each with its own uniform charge density. Although this is a situation where charge density in the full sphere is not uniform, the charge density function depends only on the distance from the center and not on the direction. Therefore, this charge distribution does have spherical symmetry.


ρ_0
ρ_1
ρ_2
(a) Spherically symmetric
(b) Not spherically symmetric
(c) Spherically symmetric

Illustrations of spherically symmetrical and nonsymmetrical systems. Different shadings indicate different charge densities. Charges on spherically shaped objects do not necessarily mean the charges are distributed with spherical symmetry. The spherical symmetry occurs only when the charge density does not depend on the direction. In (a), charges are distributed uniformly in a sphere. In (b), the upper half of the sphere has a different charge density from the lower half; therefore, (b) does not have spherical symmetry. In (c), the charges are in spherical shells of different charge densities, which means that charge density is only a function of the radial distance from the center; therefore, the system has spherical symmetry.

One good way to determine whether or not your problem has spherical symmetry is to look at the charge density function in spherical coordinates, $\rho\left(r,\theta,\phi\right)$. If the charge density is only a function of r, that is $\rho=\rho\left(r\right)$, then you have spherical symmetry. If the density depends on θ or ϕ, you could change it by rotation; hence, you would not have spherical symmetry.

Consequences of symmetry

In all spherically symmetrical cases, the electric field at any point must be radially directed, because the charge and, hence, the field must be invariant under rotation. Therefore, using spherical coordinates with their origins at the center of the spherical charge distribution, we can write down the expected form of the electric field at a point P located at a distance r from the center:

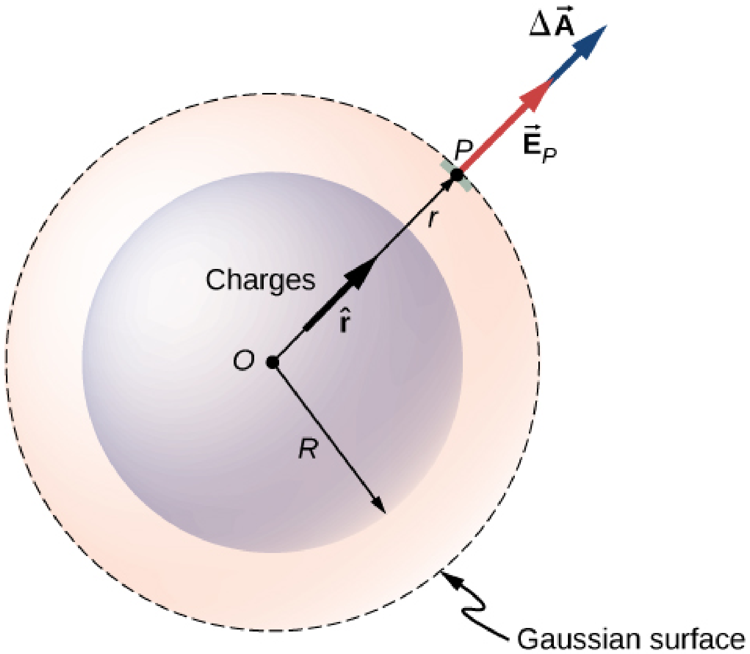
Equation:

$$\text{Spherical symmetry: }\vec{\mathbf{E}}_P=E_P(r)\hat{\mathbf{r}},$$

where $\hat{\mathbf{r}}$ is the unit vector pointed in the direction from the origin to the field point P. The radial component E_P of the electric field can be positive or negative. When $E_P>0$, the electric field at P points away from the origin, and when $E_P<0$, the electric field at P points toward the origin.

Gaussian surface and flux calculations

We can now use this form of the electric field to obtain the flux of the electric field through the Gaussian surface. For spherical symmetry, the Gaussian surface is a closed spherical surface that has the same center as the center of the charge distribution. Thus, the direction of the area vector of an area element on the Gaussian surface at any point is parallel to the direction of the electric field at that point, since they are both radially directed outward ([link]).



The electric field at any point of the spherical Gaussian surface for a spherically symmetrical charge distribution is parallel to the area element vector at that point, giving flux as the product of the magnitude of electric field and the value of the area. Note that the radius *R* of the charge distribution and the radius *r* of the Gaussian surface are different quantities.

The magnitude of the electric field $\overrightarrow{\mathbf{E}}$ must be the same everywhere on a spherical Gaussian surface concentric with the distribution. For a spherical surface of radius *r*,
Equation:

$$\Phi = \oint_S \vec{\mathbf{E}}_P \cdot \widehat{\mathbf{n}}\, dA = E_P \oint_S dA = E_P\, 4\pi r^2.$$

Using Gauss’s law

According to Gauss’s law, the flux through a closed surface is equal to the total charge enclosed within the closed surface divided by the permittivity of vacuum ε_0. Let q_{enc} be the total charge enclosed inside the distance *r* from the origin, which is the space inside the Gaussian spherical surface of radius *r*. This gives the following relation for Gauss’s law:
Equation:

$$4\pi r^2 E = \frac{q_{\text{enc}}}{\varepsilon_0}.$$

Hence, the electric field at point *P* that is a distance *r* from the center of a spherically symmetrical charge distribution has the following magnitude and direction:
Equation:

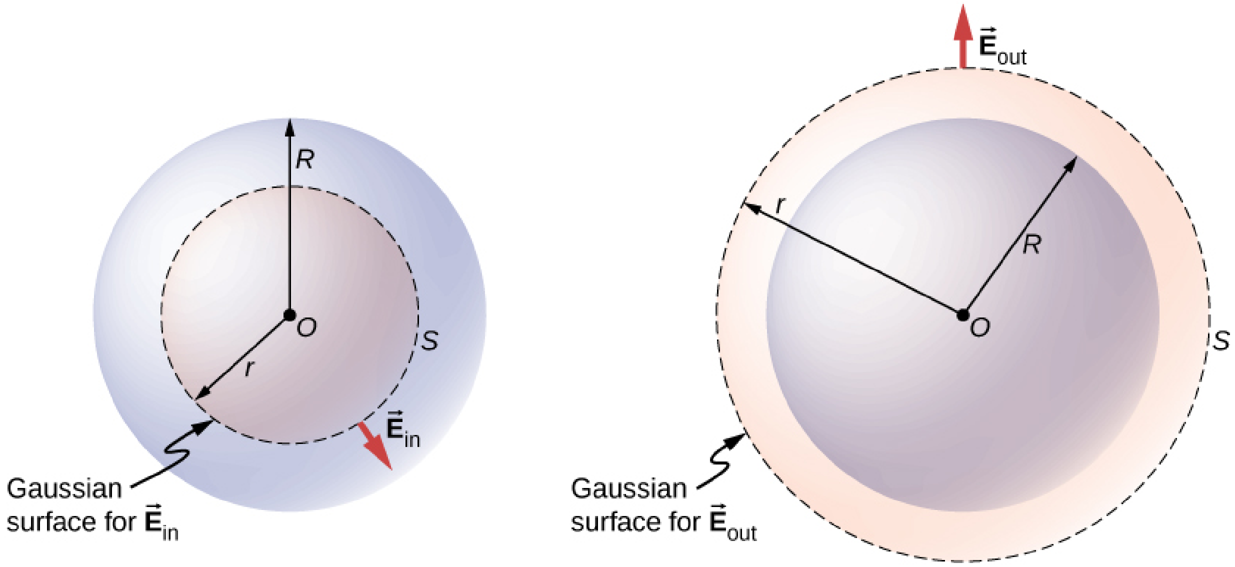
$$\text{Magnitude: } E(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_{\text{enc}}}{r^2}$$

Direction: radial from *O* to *P* or from *P* to *O*.

The direction of the field at point *P* depends on whether the charge in the sphere is positive or negative. For a net positive charge enclosed within the Gaussian surface, the direction is from *O* to *P*, and for a net negative charge, the direction is from *P* to *O*. This is all we need for a point charge, and you will notice that the result above is identical to that for a point charge. However, Gauss’s law becomes truly useful in cases where the charge occupies a finite volume.

Computing enclosed charge

The more interesting case is when a spherical charge distribution occupies a volume, and asking what the electric field inside the charge distribution is thus becomes relevant. In this case, the charge enclosed depends on the distance *r* of the field point relative to the radius of the charge distribution *R*, such as that shown in [link].


R
O
S
r
$\vec{E}_{in}$
Gaussian surface for $\vec{E}_{in}$
$\vec{E}_{out}$
r
R
O
S
Gaussian surface for $\vec{E}_{out}$

A spherically symmetrical charge distribution and the Gaussian surface used for finding the field (a) inside and (b) outside the distribution.

If point P is located outside the charge distribution—that is, if $r \geq R$—then the Gaussian surface containing P encloses all charges in the sphere. In this case, q_{enc} equals the total charge in the sphere. On the other hand, if point P is within the spherical charge distribution, that is, if $r < R$, then the Gaussian surface encloses a smaller sphere than the sphere of charge distribution. In this case, q_{enc} is less than the total charge present in the sphere. Referring to [link], we can write q_{enc} as

Equation:

$$q_{\text{enc}} = \begin{cases} q_{\text{tot}}(\text{total charge}) \text{ if } r \geq R \\ q_{\text{within } r < R}(\text{only charge within } r < R) \text{ if } r < R \end{cases}.$$

The field at a point outside the charge distribution is also called $\vec{\mathbf{E}}_{\text{out}}$, and the field at a point inside the charge distribution is called $\vec{\mathbf{E}}_{\text{in}}$. Focusing on the two types of field points, either inside or outside the charge distribution, we can now write the magnitude of the electric field as

Equation:

$$P \text{ outside sphere } E_{\text{out}} = \frac{1}{4\pi\varepsilon_0} \frac{q_{\text{tot}}}{r^2}$$

Equation:

$$P \text{ inside sphere } E_{\text{in}} = \frac{1}{4\pi\varepsilon_0} \frac{q_{\text{within } r < R}}{r^2}.$$

Note that the electric field outside a spherically symmetrical charge distribution is identical to that of a point charge at the center that has a charge equal to the total charge of the spherical charge distribution. This is remarkable since the charges are not located at the center only. We now work out specific examples of spherical charge distributions, starting with the case of a uniformly charged sphere.

Example:

Uniformly Charged Sphere

A sphere of radius *R*, such as that shown in [link], has a uniform volume charge density ρ_0. Find the electric field at a point outside the sphere and at a point inside the sphere.

Strategy

Apply the Gauss's law problem-solving strategy, where we have already worked out the flux calculation.

Solution

The charge enclosed by the Gaussian surface is given by

Equation:

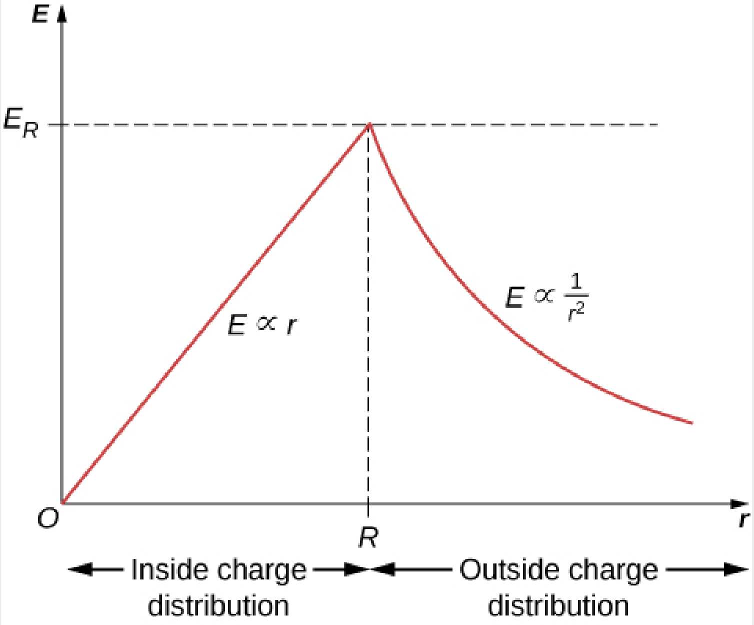
$$q_{\text{enc}} = \int_0 \rho\, dV = \int_0^r \rho_0 4\pi r'^2 dr' = \rho_0 \left(\frac{4}{3}\pi r^3\right).$$

The answer for electric field amplitude can then be written down immediately for a point outside the sphere, labeled E_{out}, and a point inside the sphere, labeled E_{in}.

Equation:

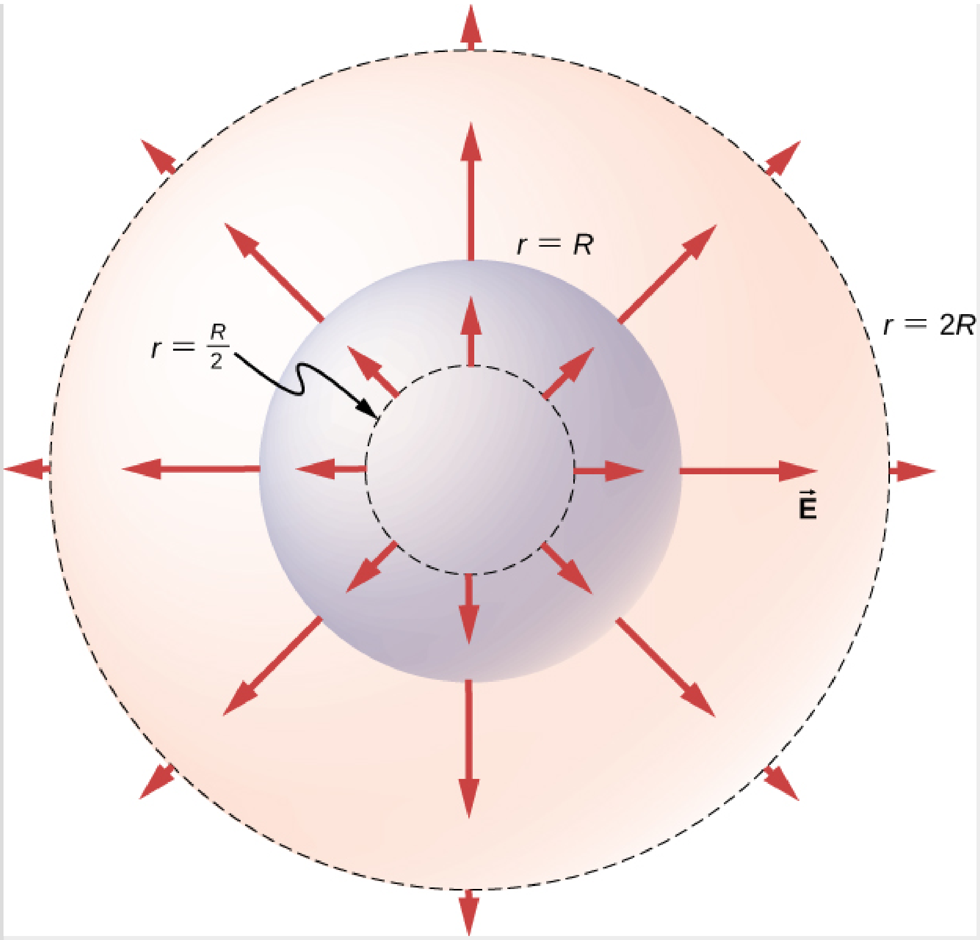
$$\begin{aligned} E_{\text{out}} &= \tfrac{1}{4\pi\varepsilon_0}\tfrac{q_{\text{tot}}}{r^2},\ q_{\text{tot}} = \tfrac{4}{3}\pi R^3\,\rho_0, \\ E_{\text{in}} &= \tfrac{q_{\text{enc}}}{4\pi\varepsilon_0 r^2} = \tfrac{\rho_0 r}{3\varepsilon_0},\ \text{since } q_{\text{enc}} = \tfrac{4}{3}\pi r^3\,\rho_0. \end{aligned}$$

It is interesting to note that the magnitude of the electric field increases inside the material as you go out, since the amount of charge enclosed by the Gaussian surface increases with the volume. Specifically, the charge enclosed grows $\propto r^3$, whereas the field from each infinitesimal element of charge drops off $\propto 1/r^2$ with the net result that the electric field within the distribution increases in strength linearly with the radius. The magnitude of the electric field outside the sphere decreases as you go away from the charges, because the included charge remains the same but the distance increases. [link] displays the variation of the magnitude of the electric field with distance from the center of a uniformly charged sphere.



Electric field of a uniformly charged, non-conducting sphere increases inside the sphere to a maximum at the surface and then decreases as $1/r^2$. Here, $E_R = \frac{\rho_0 R}{3\varepsilon_0}$. The electric field is due to a spherical charge distribution of uniform charge density and total charge Q as a function of distance from the center of the distribution.

The direction of the electric field at any point P is radially outward from the origin if ρ_0 is positive, and inward (i.e., toward the center) if ρ_0 is negative. The electric field at some representative space points are displayed in [link] whose radial coordinates r are $r = R/2$, $r = R$, and $r = 2R$.



Electric field vectors inside and outside a uniformly charged sphere.

Significance
Notice that E_{out} has the same form as the equation of the electric field of an isolated point charge. In determining the electric field of a uniform spherical charge distribution, we can therefore assume that all of the charge inside the appropriate spherical Gaussian surface is located at the center of the distribution.

Example:
Non-Uniformly Charged Sphere
A non-conducting sphere of radius R has a non-uniform charge density that varies with the distance from its center as given by
Equation:

$$\rho(r) = ar^n \; (r \le R; n \ge 0),$$

where a is a constant. We require $n \ge 0$ so that the charge density is not undefined at $r = 0$. Find the electric field at a point outside the sphere and at a point inside the sphere.

Strategy

Apply the Gauss’s law strategy given above, where we work out the enclosed charge integrals separately for cases inside and outside the sphere.

Solution

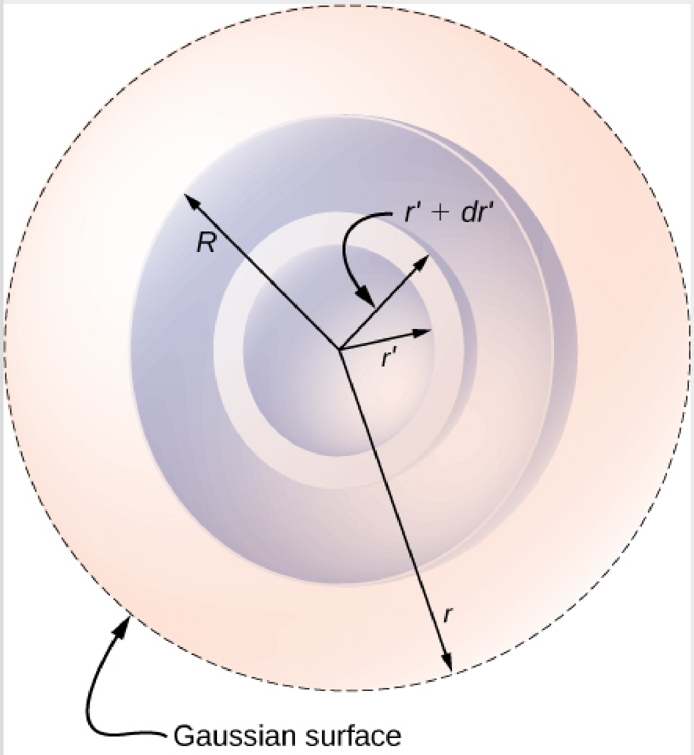
Since the given charge density function has only a radial dependence and no dependence on direction, we have a spherically symmetrical situation. Therefore, the magnitude of the electric field at any point is given above and the direction is radial. We just need to find the enclosed charge q_{enc}, which depends on the location of the field point.

A note about symbols: We use r' for locating charges in the charge distribution and r for locating the field point(s) at the Gaussian surface(s). The letter R is used for the radius of the charge distribution.

As charge density is not constant here, we need to integrate the charge density function over the volume enclosed by the Gaussian surface. Therefore, we set up the problem for charges in one spherical shell, say between r' and $r' + dr'$, as shown in [link]. The volume of charges in the shell of infinitesimal width is equal to the product of the area of surface $4\pi r'^2$ and the thickness dr'. Multiplying the volume with the density at this location, which is ar'^n, gives the charge in the shell:

Equation:

$$dq = ar'^n \, 4\pi r'^2 dr'.$$



Spherical symmetry with non-uniform charge distribution. In this type of problem, we need four radii: R is the radius of the charge distribution, r is the radius of the Gaussian surface, r' is the inner radius of the spherical shell, and $r' + dr'$ is the outer radius of the spherical shell. The spherical shell is used to calculate the charge enclosed within the Gaussian surface. The range for r' is from 0 to *r* for the field at a point inside the charge distribution and from 0 to *R* for the field at a point outside the charge distribution. If $r > R$, then the Gaussian surface encloses more volume than the charge distribution, but the additional volume does not contribute to q_{enc}.

(a) **Field at a point outside the charge distribution.** In this case, the Gaussian surface, which contains the field point *P*, has a radius *r* that is greater than the radius *R* of the charge distribution, $r > R$. Therefore, all charges of the charge distribution are enclosed within the Gaussian surface. Note that the space between $r' = R$ and $r' = r$ is empty of charges and therefore does not contribute to the integral over the volume enclosed by the Gaussian surface:

Equation:

$$q_{\text{enc}} = \int dq = \int_0^R ar'^n\, 4\pi r'^2 dr' = \frac{4\pi a}{n+3} R^{n+3}.$$

This is used in the general result for $\vec{\mathbf{E}}_{\text{out}}$ above to obtain the electric field at a point outside the charge distribution as

Equation:

$$\vec{\mathbf{E}}_{\text{out}} = \left[\frac{aR^{n+3}}{\varepsilon_0(n+3)}\right] \frac{1}{r^2}\hat{\mathbf{r}},$$

where $\hat{\mathbf{r}}$ is a unit vector in the direction from the origin to the field point at the Gaussian surface.

(b) **Field at a point inside the charge distribution.** The Gaussian surface is now buried inside the charge distribution, with $r < R$. Therefore, only those charges in the distribution that are within a distance *r* of the center of the spherical charge distribution count in r_{enc}:

Equation:

$$q_{\text{enc}} = \int_0^r ar'^n 4\pi r'^2\, dr' = \frac{4\pi a}{n+3} r^{n+3}.$$

Now, using the general result above for $\vec{\mathbf{E}}_{\text{in}}$, we find the electric field at a point that is a distance *r* from the center and lies within the charge distribution as

Equation:

$$\vec{\mathbf{E}}_{\text{in}} = \left[\frac{a}{\varepsilon_0(n+3)}\right] r^{n+1}\hat{\mathbf{r}},$$

where the direction information is included by using the unit radial vector.

Note:
Exercise:

Problem:

Check Your Understanding Check that the electric fields for the sphere reduce to the correct values for a point charge.

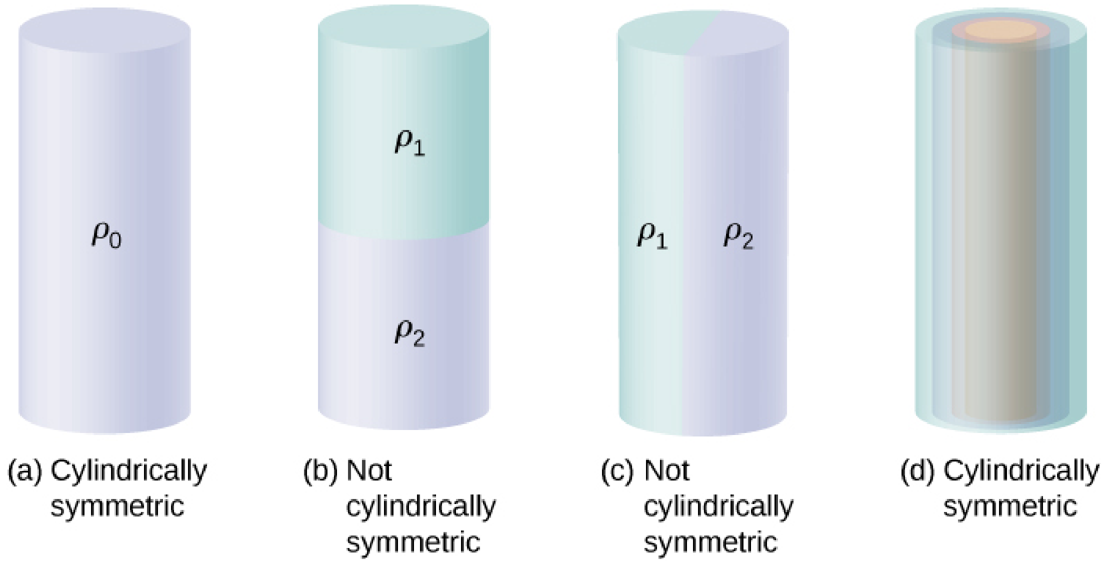
Solution:

In this case, there is only $\vec{\mathbf{E}}_{\text{out}}.$ So, yes.

Charge Distribution with Cylindrical Symmetry

A charge distribution has **cylindrical symmetry** if the charge density depends only upon the distance *r* from the axis of a cylinder and must not vary along the axis or with direction about the axis. In other words, if your system varies if you rotate it around the axis, or shift it along the axis, you do not have cylindrical symmetry.

[link] shows four situations in which charges are distributed in a cylinder. A uniform charge density $\rho_0.$ in an infinite straight wire has a cylindrical symmetry, and so does an infinitely long cylinder with constant charge density $\rho_0.$ An infinitely long cylinder that has different charge densities along its length, such as a charge density ρ_1 for $z > 0$ and $\rho_2 \neq \rho_1$ for $z < 0$, does not have a usable cylindrical symmetry for this course. Neither does a cylinder in which charge density varies with the direction, such as a charge density ρ_1 for $0 \leq \theta < \pi$ and $\rho_2 \neq \rho_1$ for $\pi \leq \theta < 2\pi$. A system with concentric cylindrical shells, each with uniform charge densities, albeit different in different shells, as in [link](d), does have cylindrical symmetry if they are infinitely long. The infinite length requirement is due to the charge density changing along the axis of a finite cylinder. In real systems, we don’t have infinite cylinders; however, if the cylindrical object is considerably longer than the radius from it that we are interested in, then the approximation of an infinite cylinder becomes useful.



ρ_0
ρ_1
ρ_2
ρ_1
ρ_2
(a) Cylindrically symmetric
(b) Not cylindrically symmetric
(c) Not cylindrically symmetric
(d) Cylindrically symmetric

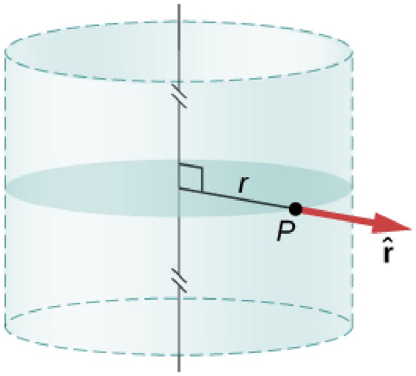
To determine whether a given charge distribution has cylindrical symmetry, look at the cross-section of an “infinitely long” cylinder. If the charge density does not depend on the polar angle of the cross-section or along the axis, then you have cylindrical symmetry. (a) Charge density is constant in the cylinder; (b) upper half of the cylinder has a different charge density from the lower half; (c) left half of the cylinder has a different charge density from the right half; (d) charges are constant in different cylindrical rings, but the density does not depend on the polar angle. Cases (a) and (d) have cylindrical symmetry, whereas (b) and (c) do not.

Consequences of symmetry

In all cylindrically symmetrical cases, the electric field $\overrightarrow{\mathbf{E}}_P$ at any point *P* must also display cylindrical symmetry.

Cylindrical symmetry: $\vec{\mathbf{E}}_P = E_P(r)\hat{\mathbf{r}}$,

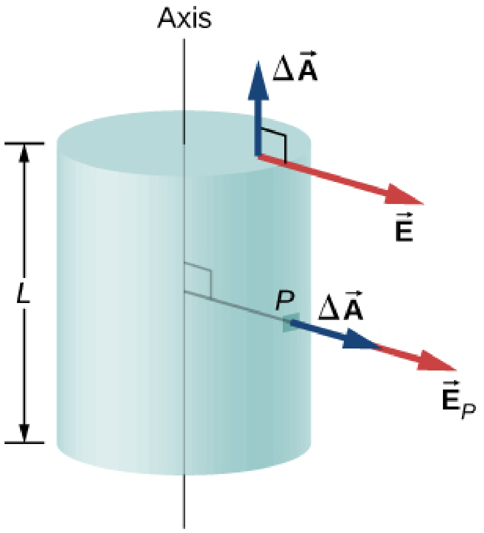
where *r* is the distance from the axis and $\hat{\mathbf{r}}$ is a unit vector directed perpendicularly away from the axis ([link]).



The electric field in a cylindrically symmetrical situation depends only on the distance from the axis. The direction of the electric field is pointed away from the axis for positive charges and toward the axis for negative charges.

Gaussian surface and flux calculation

To make use of the direction and functional dependence of the electric field, we choose a closed Gaussian surface in the shape of a cylinder with the same axis as the axis of the charge distribution. The flux through this surface of radius *s* and height *L* is easy to compute if we divide our task into two parts: (a) a flux through the flat ends and (b) a flux through the curved surface ([link]).



The Gaussian surface in the case of cylindrical symmetry. The electric field at a patch is either parallel or perpendicular to the normal to the patch of the Gaussian surface.

The electric field is perpendicular to the cylindrical side and parallel to the planar end caps of the surface. The flux through the cylindrical part is

Equation:

$$\int_S \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = E \int_S dA = E(2\pi r L),$$

whereas the flux through the end caps is zero because $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}} = 0$ there. Thus, the flux is

Equation:

$$\int_S \overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = E(2\pi r L) + 0 + 0 = 2\pi r L E.$$

Using Gauss’s law

According to Gauss’s law, the flux must equal the amount of charge within the volume enclosed by this surface, divided by the permittivity of free space. When you do the calculation for a cylinder of length L, you find that q_{enc} of Gauss’s law is directly proportional to L. Let us write it as charge per unit length (λ_{enc}) times length L:

Equation:

$$q_{\text{enc}} = \lambda_{\text{enc}} L.$$

Hence, Gauss's law for any cylindrically symmetrical charge distribution yields the following magnitude of the electric field a distance *s* away from the axis:

Equation:

$$\text{Magnitude: } E(r) = \frac{\lambda_{\text{enc}}}{2\pi\varepsilon_0}\frac{1}{r}.$$

The charge per unit length λ_{enc} depends on whether the field point is inside or outside the cylinder of charge distribution, just as we have seen for the spherical distribution.

Computing enclosed charge

Let *R* be the radius of the cylinder within which charges are distributed in a cylindrically symmetrical way. Let the field point *P* be at a distance *s* from the axis. (The side of the Gaussian surface includes the field point *P*.) When $r > R$ (that is, when *P* is outside the charge distribution), the Gaussian surface includes all the charge in the cylinder of radius *R* and length *L*. When $r < R$ (*P* is located inside the charge distribution), then only the charge within a cylinder of radius *s* and length *L* is enclosed by the Gaussian surface:

Equation:

$$\lambda_{\text{enc}} L = \begin{cases} (\text{total charge}) \text{ if } r \geq R \\ (\text{only charge within } r < R) \text{ if } r < R \end{cases}.$$

Example:

Uniformly Charged Cylindrical Shell

A very long non-conducting cylindrical shell of radius *R* has a uniform surface charge density σ_0. Find the electric field (a) at a point outside the shell and (b) at a point inside the shell.

Strategy

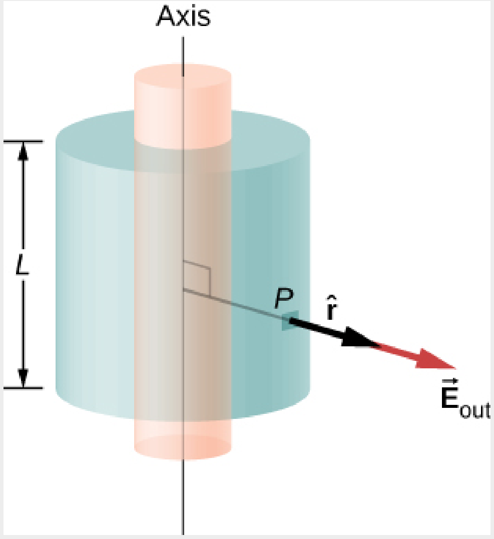
Apply the Gauss's law strategy given earlier, where we treat the cases inside and outside the shell separately.

Solution

a. **Electric field at a point outside the shell.** For a point outside the cylindrical shell, the Gaussian surface is the surface of a cylinder of radius $r > R$ and length *L*, as shown in [link]. The charge enclosed by the Gaussian cylinder is equal to the charge on the cylindrical shell of length *L*. Therefore, λ_{enc} is given by

Equation:

$$\lambda_{\text{enc}} = \frac{\sigma_0 2\pi\, RL}{L} = 2\pi\, R\sigma_0.$$



A Gaussian surface surrounding a cylindrical shell.

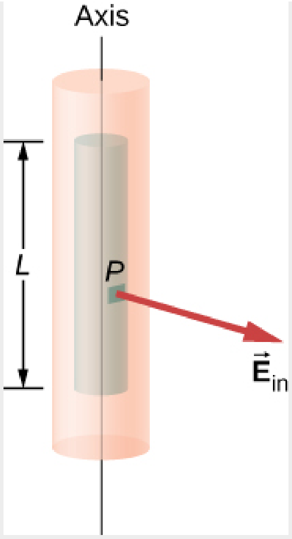
Hence, the electric field at a point *P* outside the shell at a distance *r* away from the axis is
Equation:

$$\vec{\mathbf{E}} = \frac{2\pi R\sigma_0}{2\pi\varepsilon_o}\,\frac{1}{r}\,\hat{\mathbf{r}} = \frac{R\sigma_0}{\varepsilon_o}\,\frac{1}{r}\,\hat{\mathbf{r}}\;(r > R)$$

where $\hat{\mathbf{r}}$ is a unit vector, perpendicular to the axis and pointing away from it, as shown in the figure. The electric field at *P* points in the direction of $\hat{\mathbf{r}}$ given in [link] if $\sigma_0 > 0$ and in the opposite direction to $\hat{\mathbf{r}}$ if $\sigma_0 < 0$.

b. **Electric field at a point inside the shell.** For a point inside the cylindrical shell, the Gaussian surface is a cylinder whose radius *r* is less than *R* ([link]). This means no charges are included inside the Gaussian surface:
Equation:

$$\lambda_{\text{enc}} = 0.$$



A Gaussian surface within a cylindrical shell.

This gives the following equation for the magnitude of the electric field E_{in} at a point whose *r* is less than *R* of the shell of charges.

Equation:

$$E_{\text{in}}\,2\pi rL = 0\,(r < R),$$

This gives us

Equation:

$$E_{\text{in}} = 0\,(r < R).$$

Significance

Notice that the result inside the shell is exactly what we should expect: No enclosed charge means zero electric field. Outside the shell, the result becomes identical to a wire with uniform charge $R\sigma_0$.

Note:

Exercise:

Problem:

Check Your Understanding A thin straight wire has a uniform linear charge density λ_0. Find the electric field at a distance *d* from the wire, where *d* is much less than the length of the wire.

Solution:

$\overrightarrow{\mathbf{E}} = \frac{\lambda_0}{2\pi\varepsilon_0}\,\frac{1}{d}\hat{\mathbf{r}}$; This agrees with the calculation of [link] where we found the electric field by integrating over the charged wire. Notice how much simpler the calculation of this electric field is with Gauss's law.

Charge Distribution with Planar Symmetry

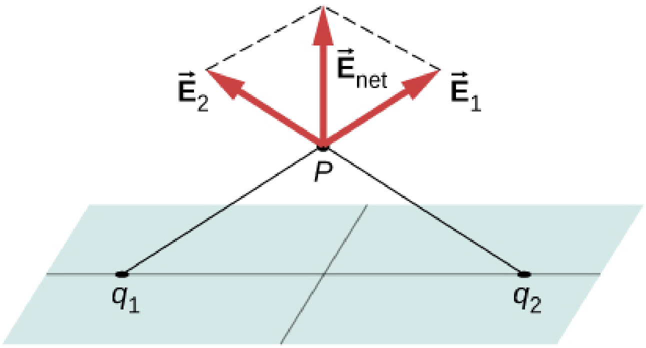
A **planar symmetry** of charge density is obtained when charges are uniformly spread over a large flat surface. In planar symmetry, all points in a plane parallel to the plane of charge are identical with respect to the charges.

Consequences of symmetry

We take the plane of the charge distribution to be the *xy*-plane and we find the electric field at a space point *P* with coordinates (*x*, *y*, *z*). Since the charge density is the same at all (*x*, *y*)-coordinates in the $z = 0$ plane, by symmetry, the electric field at *P* cannot depend on the *x*- or *y*-coordinates of point *P*, as shown in [link]. Therefore, the electric field at *P* can only depend on the distance from the plane and has a direction either toward the plane or away from the plane. That is, the electric field at *P* has only a nonzero *z*-component.

Uniform charges in *xy* plane: $\overrightarrow{\mathbf{E}} = E(z)\widehat{\mathbf{z}}$

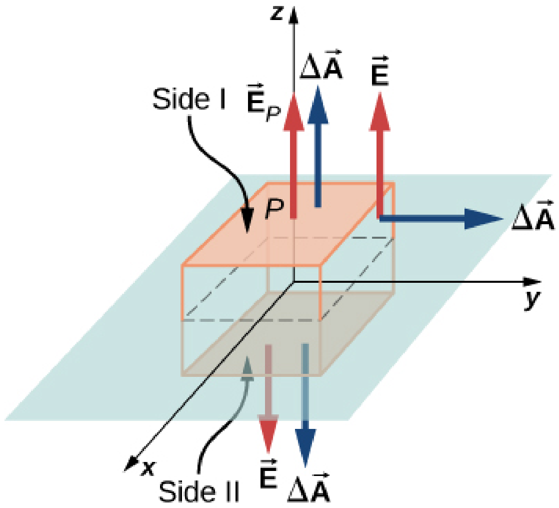
where *z* is the distance from the plane and $\widehat{\mathbf{z}}$ is the unit vector normal to the plane. Note that in this system, $E\left(z\right) = E\left(-z\right)$, although of course they point in opposite directions.



The components of the electric field parallel to a plane of charges cancel out the two charges located symmetrically from the field point P. Therefore, the field at any point is pointed vertically from the plane of charges. For any point P and charge q_1, we can always find a q_2 with this effect.

Gaussian surface and flux calculation

In the present case, a convenient Gaussian surface is a box, since the expected electric field points in one direction only. To keep the Gaussian box symmetrical about the plane of charges, we take it to straddle the plane of the charges, such that one face containing the field point P is taken parallel to the plane of the charges. In [link], sides I and II of the Gaussian surface (the box) that are parallel to the infinite plane have been shaded. They are the only surfaces that give rise to nonzero flux because the electric field and the area vectors of the other faces are perpendicular to each other.



A thin charged sheet and the Gaussian box for finding the electric field at the field point *P*. The normal to each face of the box is from inside the box to outside. On two faces of the box, the electric fields are parallel to the area vectors, and on the other four faces, the electric fields are perpendicular to the area vectors.

Let *A* be the area of the shaded surface on each side of the plane and E_P be the magnitude of the electric field at point *P*. Since sides I and II are at the same distance from the plane, the electric field has the same magnitude at points in these planes, although the directions of the electric field at these points in the two planes are opposite to each other.

Magnitude at I or II: $E(z) = E_P$.

If the charge on the plane is positive, then the direction of the electric field and the area vectors are as shown in [link]. Therefore, we find for the flux of electric field through the box

Equation:

$$\Phi = \oint_S \overrightarrow{\mathbf{E}}_P \cdot \widehat{\mathbf{n}}\, dA = E_P A + E_P A + 0 + 0 + 0 + 0 = 2E_P A$$

where the zeros are for the flux through the other sides of the box. Note that if the charge on the plane is negative, the directions of electric field and area vectors for planes I and II are opposite to each other, and we get a negative sign for the flux. According to Gauss's law, the flux must equal $q_{\text{enc}}/\varepsilon_0$. From [link], we see that the charges inside the volume enclosed by the Gaussian box reside on an area *A* of the *xy*-plane. Hence,

Equation:

$$q_{\text{enc}} = \sigma_0 A.$$

Using the equations for the flux and enclosed charge in Gauss's law, we can immediately determine the electric field at a point at height *z* from a uniformly charged plane in the *xy*-plane:
Equation:

$$\vec{\mathbf{E}}_P = \frac{\sigma_0}{2\varepsilon_0}\widehat{\mathbf{n}}.$$

The direction of the field depends on the sign of the charge on the plane and the side of the plane where the field point *P* is located. Note that above the plane, $\widehat{\mathbf{n}} = +\widehat{\mathbf{z}}$, while below the plane, $\widehat{\mathbf{n}} = -\widehat{\mathbf{z}}$.

You may be surprised to note that the electric field does not actually depend on the distance from the plane; this is an effect of the assumption that the plane is infinite. In practical terms, the result given above is still a useful approximation for finite planes near the center.

Summary

- For a charge distribution with certain spatial symmetries (spherical, cylindrical, and planar), we can find a Gaussian surface over which $\overrightarrow{\mathbf{E}} \cdot \widehat{\mathbf{n}} = E$, where E is constant over the surface. The electric field is then determined with Gauss's law.
- For spherical symmetry, the Gaussian surface is also a sphere, and Gauss's law simplifies to $4\pi r^2 E = \frac{q_{\text{enc}}}{\varepsilon_0}$.
- For cylindrical symmetry, we use a cylindrical Gaussian surface, and find that Gauss's law simplifies to $2\pi r L E = \frac{q_{\text{enc}}}{\varepsilon_0}$.
- For planar symmetry, a convenient Gaussian surface is a box penetrating the plane, with two faces parallel to the plane and the remainder perpendicular, resulting in Gauss's law being $2AE = \frac{q_{\text{enc}}}{\varepsilon_0}$.

Conceptual Questions

Exercise:

Problem:

Would Gauss's law be helpful for determining the electric field of two equal but opposite charges a fixed distance apart?

Solution:

No, since the situation does not have symmetry, making Gauss's law challenging to simplify.

Exercise:

Problem:

Discuss the role that symmetry plays in the application of Gauss’s law. Give examples of continuous charge distributions in which Gauss’s law is useful and not useful in determining the electric field.

Exercise:

Problem:

Discuss the restrictions on the Gaussian surface used to discuss planar symmetry. For example, is its length important? Does the cross-section have to be square? Must the end faces be on opposite sides of the sheet?

Solution:

Any shape of the Gaussian surface can be used. The only restriction is that the Gaussian integral must be calculable; therefore, a box or a cylinder are the most convenient geometrical shapes for the Gaussian surface.

Problems

Exercise:

Problem:

Recall that in the example of a uniform charged sphere, $\rho_0 = Q/\left(\frac{4}{3}\pi R^3\right)$. Rewrite the answers in terms of the total charge Q on the sphere.

Solution:

$r > R,\ E = \frac{Q}{4\pi\varepsilon_0 r^2};\ r < R,\ E = \frac{qr}{4\pi\varepsilon_0 R^3}$

Exercise:

Problem:

Suppose that the charge density of the spherical charge distribution shown in [link] is $\rho(r) = \rho_0 r/R$ for $r \leq R$ and zero for $r > R$. Obtain expressions for the electric field both inside and outside the distribution.

Exercise:

Problem:

A very long, thin wire has a uniform linear charge density of 50 $\mu\mathrm{C/m}$. What is the electric field at a distance 2.0 cm from the wire?

Solution:

$EA = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = 4.50 \times 10^7 \text{ N/C}$

Exercise:

Problem:

A charge of $-30\ \mu\text{C}$ is distributed uniformly throughout a spherical volume of radius 10.0 cm. Determine the electric field due to this charge at a distance of (a) 2.0 cm, (b) 5.0 cm, and (c) 20.0 cm from the center of the sphere.

Exercise:

Problem:

Repeat your calculations for the preceding problem, given that the charge is distributed uniformly over the surface of a spherical conductor of radius 10.0 cm.

Solution:

a. 0; b. 0; c. $\vec{\mathbf{E}} = 6.74 \times 10^6 \text{ N/C}(-\hat{\mathbf{r}})$

Exercise:

Problem:

A total charge Q is distributed uniformly throughout a spherical shell of inner and outer radii r_1 and r_2, respectively. Show that the electric field due to the charge is

$$\begin{aligned} &\vec{\mathbf{E}} = \vec{\mathbf{0}} && (r \le r_1); \\ &\vec{\mathbf{E}} = \frac{Q}{4\pi\varepsilon_0 r^2}\left(\frac{r^3 - {r_1}^3}{{r_2}^3 - {r_1}^3}\right)\hat{\mathbf{r}} && (r_1 \le r \le r_2); \\ &\vec{\mathbf{E}} = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\mathbf{r}} && (r \ge r_2). \end{aligned}$$

Exercise:

Problem:

When a charge is placed on a metal sphere, it ends up in equilibrium at the outer surface. Use this information to determine the electric field of $+3.0\ \mu\text{C}$ charge put on a 5.0-cm aluminum spherical ball at the following two points in space: (a) a point 1.0 cm from the center of the ball (an inside point) and (b) a point 10 cm from the center of the ball (an outside point).

Solution:

a. 0; b. $E = 2.70 \times 10^6 \text{ N/C}$

Exercise:

Problem:

A large sheet of charge has a uniform charge density of 10 $\mu\mathrm{C/m^2}$. What is the electric field due to this charge at a point just above the surface of the sheet?

Exercise:

Problem:

Determine if approximate cylindrical symmetry holds for the following situations. State why or why not. (a) A 300-cm long copper rod of radius 1 cm is charged with +500 nC of charge and we seek electric field at a point 5 cm from the center of the rod. (b) A 10-cm long copper rod of radius 1 cm is charged with +500 nC of charge and we seek electric field at a point 5 cm from the center of the rod. (c) A 150-cm wooden rod is glued to a 150-cm plastic rod to make a 300-cm long rod, which is then painted with a charged paint so that one obtains a uniform charge density. The radius of each rod is 1 cm, and we seek an electric field at a point that is 4 cm from the center of the rod. (d) Same rod as (c), but we seek electric field at a point that is 500 cm from the center of the rod.

Solution:

a. Yes, the length of the rod is much greater than the distance to the point in question. b. No, The length of the rod is of the same order of magnitude as the distance to the point in question. c. Yes, the length of the rod is much greater than the distance to the point in question. d. No. The length of the rod is of the same order of magnitude as the distance to the point in question.

Exercise:

Problem:

A long silver rod of radius 3 cm has a charge of $-5\ \mu\mathrm{C/cm}$ on its surface. (a) Find the electric field at a point 5 cm from the center of the rod (an outside point). (b) Find the electric field at a point 2 cm from the center of the rod (an inside point).

Exercise:

Problem:

The electric field at 2 cm from the center of long copper rod of radius 1 cm has a magnitude 3 N/C and directed outward from the axis of the rod. (a) How much charge per unit length exists on the copper rod? (b) What would be the electric flux through a cube of side 5 cm situated such that the rod passes through opposite sides of the cube perpendicularly?

Solution:

a. $\overrightarrow{\mathbf{E}} = \frac{R\sigma_0}{\varepsilon_0}\ \frac{1}{r}\hat{\mathbf{r}} \Rightarrow \sigma_0 = 5.31\ \times\ 10^{-11}\ \mathrm{C/m^2}$,
$\lambda = 3.33\ \times\ 10^{-12}\ \mathrm{C/m}$;
b. $\Phi = \frac{q_{\text{enc}}}{\varepsilon_0} = \frac{3.33 \times 10^{-12}\ \mathrm{C/m}(0.05\ \mathrm{m})}{\varepsilon_0} = 0.019\ \mathrm{N \cdot m^2/C}$

Exercise:

Problem:

A long copper cylindrical shell of inner radius 2 cm and outer radius 3 cm surrounds concentrically a charged long aluminum rod of radius 1 cm with a charge density of 4 pC/m. All charges on the aluminum rod reside at its surface. The inner surface of the copper shell has exactly opposite charge to that of the aluminum rod while the outer surface of the copper shell has the same charge as the aluminum rod. Find the magnitude and direction of the electric field at points that are at the following distances from the center of the aluminum rod: (a) 0.5 cm, (b) 1.5 cm, (c) 2.5 cm, (d) 3.5 cm, and (e) 7 cm.

Exercise:

Problem:

Charge is distributed uniformly with a density ρ throughout an infinitely long cylindrical volume of radius *R*. Show that the field of this charge distribution is directed radially with respect to the cylinder and that

$$E = \frac{\rho r}{2\varepsilon_0} \qquad (r \leq R);$$
$$E = \frac{\rho R^2}{2\varepsilon_0 r} \qquad (r \geq R).$$

Solution:

$$E2\pi rl = \frac{\rho\pi r^2 l}{\varepsilon_0} \Rightarrow E = \frac{\rho r}{2\varepsilon_0}\ (r \leq R);$$
$$E2\pi rl = \frac{\rho\pi R^2 l}{\varepsilon_0} \Rightarrow E = \frac{\rho R^2}{2\varepsilon_0 r}\ (r \geq R)$$

Exercise:

Problem:

Charge is distributed throughout a very long cylindrical volume of radius *R* such that the charge density increases with the distance *r* from the central axis of the cylinder according to $\rho = \alpha r$, where α is a constant. Show that the field of this charge distribution is directed radially with respect to the cylinder and that

$$E = \frac{\alpha r^2}{3\varepsilon_0} \qquad (r \leq R);$$
$$E = \frac{\alpha R^3}{3\varepsilon_0 r} \qquad (r \geq R).$$

Exercise:

Problem:

The electric field 10.0 cm from the surface of a copper ball of radius 5.0 cm is directed toward the ball’s center and has magnitude $4.0\ \times\ 10^2\ \mathrm{N/C}$. How much charge is on the surface of the ball?

Solution:

$\Phi = \frac{q_{\text{enc}}}{\varepsilon_0} \Rightarrow q_{\text{enc}} = -1.0 \times 10^{-9}\ \text{C}$

Exercise:

Problem:

Charge is distributed throughout a spherical shell of inner radius r_1 and outer radius r_2 with a volume density given by $\rho = \rho_0 r_1/r$, where ρ_0 is a constant. Determine the electric field due to this charge as a function of *r*, the distance from the center of the shell.

Exercise:

Problem:

Charge is distributed throughout a spherical volume of radius *R* with a density $\rho = \alpha r^2$, where α is a constant. Determine the electric field due to the charge at points both inside and outside the sphere.

Solution:

$$q_{\text{enc}} = \frac{4}{5}\pi\alpha r^5,$$
$$E 4\pi r^2 = \frac{4\pi\alpha r^5}{5\varepsilon_0} \Rightarrow E = \frac{\alpha r^3}{5\varepsilon_0}\ (r \le R),$$
$$q_{\text{enc}} = \frac{4}{5}\pi\alpha R^5,\ E 4\pi r^2 = \frac{4\pi\alpha R^5}{5\varepsilon_0} \Rightarrow E = \frac{\alpha R^5}{5\varepsilon_0 r^2}\ (r \ge R)$$

Exercise:

Problem:

Consider a uranium nucleus to be sphere of radius $R = 7.4 \times 10^{-15}$ m with a charge of 92*e* distributed uniformly throughout its volume. (a) What is the electric force exerted on an electron when it is 3.0×10^{-15}m from the center of the nucleus? (b) What is the acceleration of the electron at this point?

Exercise:

Problem:

The volume charge density of a spherical charge distribution is given by $\rho(r) = \rho_0 e^{-\alpha r}$, where ρ_0 and α are constants. What is the electric field produced by this charge distribution?

Solution:

integrate by parts:

$$q_{\text{enc}} = 4\pi\rho_0 \left[-e^{-\alpha r}\left(\frac{(r)^2}{\alpha} + \frac{2r}{\alpha^2} + \frac{2}{\alpha^3}\right) + \frac{2}{\alpha^3}\right] \Rightarrow E = \frac{\rho_0}{r^2\varepsilon_0}\left[-e^{-\alpha r}\left(\frac{(r)^2}{\alpha} + \frac{2r}{\alpha^2} + \frac{2}{\alpha^3}\right) + \frac{2}{\alpha^3}\right]$$

Glossary

cylindrical symmetry
system only varies with distance from the axis, not direction

planar symmetry
system only varies with distance from a plane

spherical symmetry
system only varies with the distance from the origin, not in direction

Conductors in Electrostatic Equilibrium
By the end of this section, you will be able to:

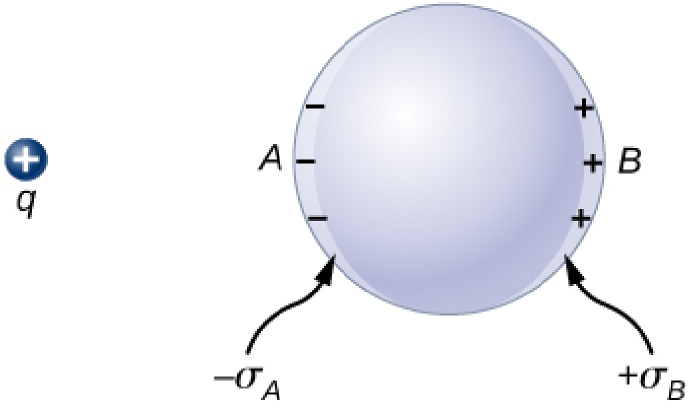
- Describe the electric field within a conductor at equilibrium
- Describe the electric field immediately outside the surface of a charged conductor at equilibrium
- Explain why if the field is not as described in the first two objectives, the conductor is not at equilibrium

So far, we have generally been working with charges occupying a volume within an insulator. We now study what happens when free charges are placed on a conductor. Generally, in the presence of a (generally external) electric field, the free charge in a conductor redistributes and very quickly reaches electrostatic equilibrium. The resulting charge distribution and its electric field have many interesting properties, which we can investigate with the help of Gauss’s law and the concept of electric potential.

The Electric Field inside a Conductor Vanishes

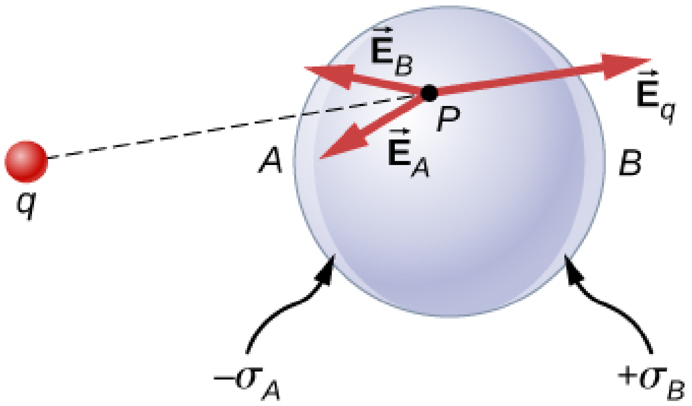
If an electric field is present inside a conductor, it exerts forces on the **free electrons** (also called conduction electrons), which are electrons in the material that are not bound to an atom. These free electrons then accelerate. However, moving charges by definition means nonstatic conditions, contrary to our assumption. Therefore, when electrostatic equilibrium is reached, the charge is distributed in such a way that the electric field inside the conductor vanishes.

If you place a piece of a metal near a positive charge, the free electrons in the metal are attracted to the external positive charge and migrate freely toward that region. The region the electrons move to then has an excess of electrons over the protons in the atoms and the region from where the electrons have migrated has more protons than electrons. Consequently, the metal develops a negative region near the charge and a positive region at the far end ([link]). As we saw in the preceding chapter, this separation of equal magnitude and opposite type of electric charge is called polarization. If you remove the external charge, the electrons migrate back and neutralize the positive region.



Polarization of a metallic sphere by an external point charge $+q$. The near side of the metal has an opposite surface charge compared to the far side of the metal. The sphere is said to be polarized. When you remove the external charge, the polarization of the metal also disappears.

The polarization of the metal happens only in the presence of external charges. You can think of this in terms of electric fields. The external charge creates an external electric field. When the metal is placed in the region of this electric field, the electrons and protons of the metal experience electric forces due to this external electric field, but only the conduction electrons are free to move in the metal over macroscopic distances. The movement of the conduction electrons leads to the polarization, which creates an induced electric field in addition to the external electric field ([link]). The net electric field is a vector sum of the fields of $+q$ and the surface charge densities $-\sigma_A$ and $+\sigma_B$. This means that the net field inside the conductor is different from the field outside the conductor.



In the presence of an external charge q, the charges in a metal redistribute. The electric field at any point has three contributions, from $+q$ and the induced charges $-\sigma_A$ and $+\sigma_B$. Note that the surface charge distribution will not be uniform in this case.

The redistribution of charges is such that the sum of the three contributions at any point P inside the conductor is

Equation:

$$\vec{\mathbf{E}}_P = \vec{\mathbf{E}}_q + \vec{\mathbf{E}}_B + \vec{\mathbf{E}}_A = \vec{\mathbf{0}}.$$

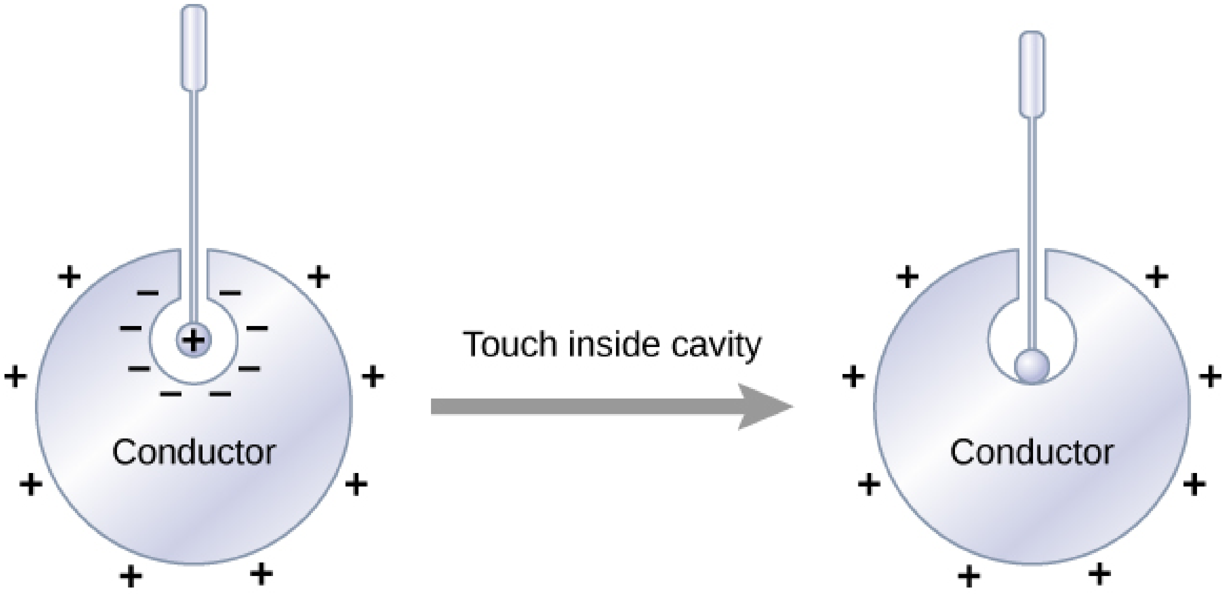
Now, thanks to Gauss's law, we know that there is no net charge enclosed by a Gaussian surface that is solely within the volume of the conductor at equilibrium. That is, $q_{\text{enc}} = 0$ and hence

Equation:

$$\vec{\mathbf{E}}_{\text{net}} = \vec{\mathbf{0}} \text{ (at points inside a conductor).}$$

Charge on a Conductor

An interesting property of a conductor in static equilibrium is that extra charges on the conductor end up on the outer surface of the conductor, regardless of where they originate. [link] illustrates a system in which we bring an external positive charge inside the cavity of a metal and then touch it to the inside surface. Initially, the inside surface of the cavity is negatively charged and the outside surface of the conductor is positively charged. When we touch the inside surface of the cavity, the induced charge is neutralized, leaving the outside surface and the whole metal charged with a net positive charge.



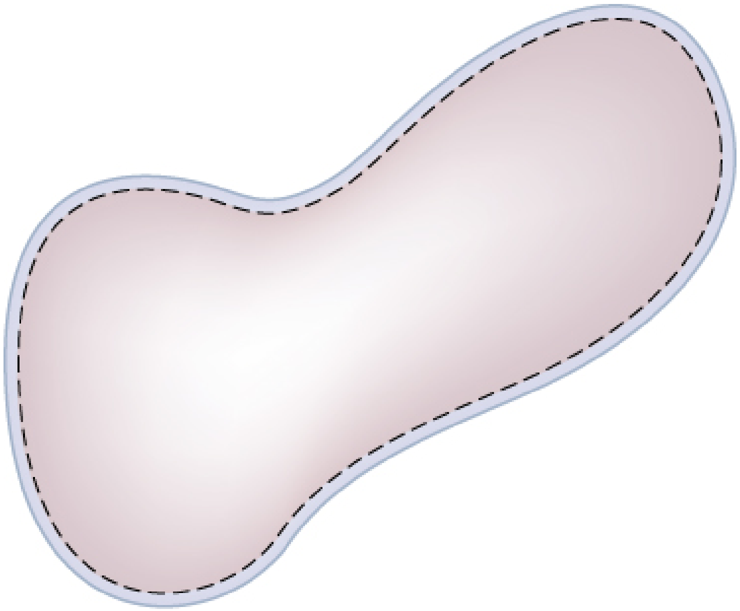
Electric charges on a conductor migrate to the outside surface no matter where you put them initially.

To see why this happens, note that the Gaussian surface in [link] (the dashed line) follows the contour of the actual surface of the conductor and is located an infinitesimal distance *within* it. Since $E = 0$ everywhere inside a conductor,

Equation:

$$\oint_{s} \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}} dA = 0.$$

Thus, from Gauss’s law, there is no net charge inside the Gaussian surface. But the Gaussian surface lies just below the actual surface of the conductor; consequently, there is no net charge inside the conductor. Any excess charge must lie on its surface.



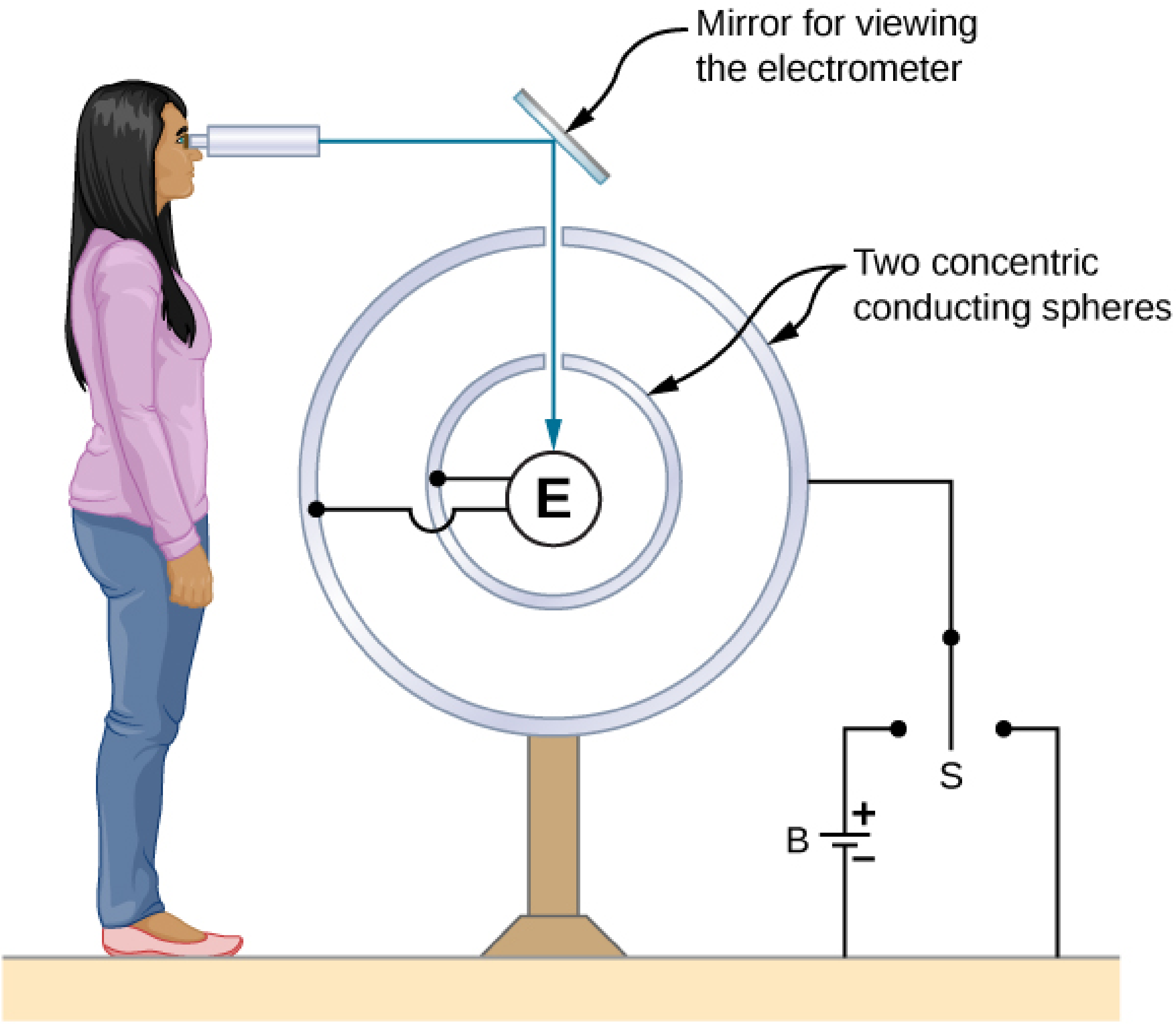
The dashed line represents a Gaussian surface that is just beneath the actual surface of the conductor.

This particular property of conductors is the basis for an extremely accurate method developed by Plimpton and Lawton in 1936 to verify Gauss’s law and, correspondingly, Coulomb’s law. A sketch of their apparatus is shown in [link]. Two spherical shells are connected to one another through an electrometer E, a device that can detect a very slight amount of charge flowing from one shell to the other. When switch S is thrown to the left, charge is placed on the outer shell by the battery B. Will charge flow through the electrometer to the inner shell?

No. Doing so would mean a violation of Gauss's law. Plimpton and Lawton did not detect any flow and, knowing the sensitivity of their electrometer, concluded that if the radial dependence in Coulomb's law were $1/r^{(2+\delta)}$, δ would be less than 2×10^{-9}[footnote]. More recent measurements place δ at less than 3×10^{-16}[footnote], a number so small that the validity of Coulomb's law seems indisputable.

S. Plimpton and W. Lawton. 1936. "A Very Accurate Test of Coulomb's Law of Force between Charges." *Physical Review* 50, No. 11: 1066, doi:10.1103/PhysRev.50.1066

E. Williams, J. Faller, and H. Hill. 1971. "New Experimental Test of Coulomb's Law: A Laboratory Upper Limit on the Photon Rest Mass." *Physical Review Letters* 26 , No. 12: 721, doi:10.1103/PhysRevLett.26.721


Mirror for viewing the electrometer
Two concentric conducting spheres
E
S
B
+
–

A representation of the apparatus used by Plimpton and Lawton. Any transfer of charge between the spheres is detected by the electrometer E.

The Electric Field at the Surface of a Conductor

If the electric field had a component parallel to the surface of a conductor, free charges on the surface would move, a situation contrary to the assumption of electrostatic equilibrium. Therefore, the electric field is always perpendicular to the surface of a conductor.

At any point just above the surface of a conductor, the surface charge density σ and the magnitude of the electric field E are related by

Note:
Equation:

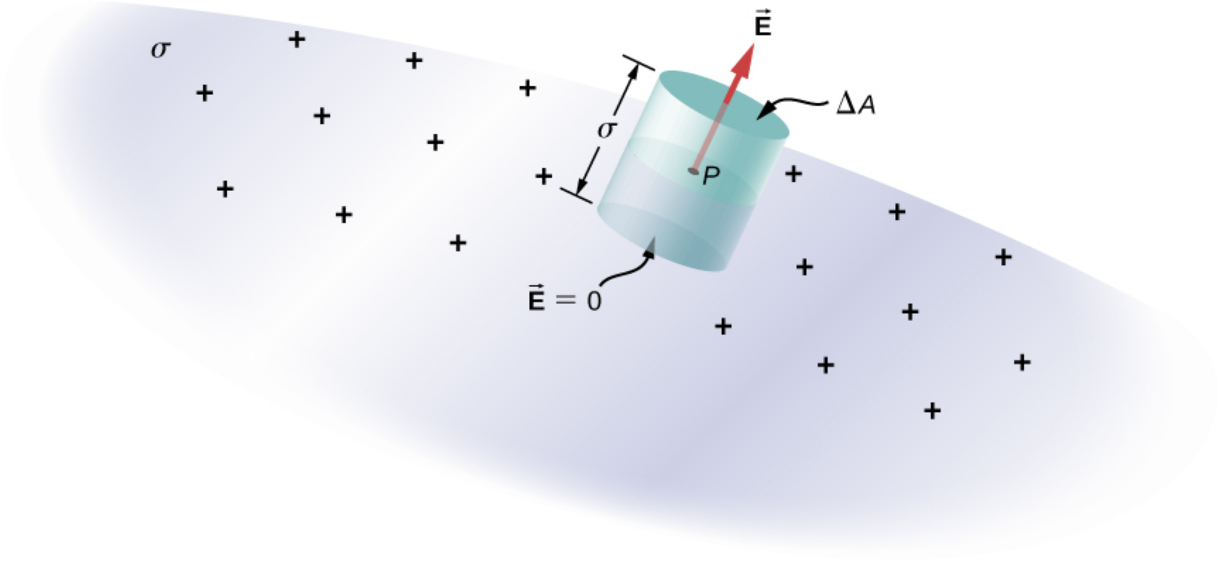
$$E = \frac{\sigma}{\varepsilon_0}.$$

To see this, consider an infinitesimally small Gaussian cylinder that surrounds a point on the surface of the conductor, as in [link]. The cylinder has one end face inside and one end face outside the surface. The height and cross-sectional area of the cylinder are δ and ΔA, respectively. The cylinder's sides are perpendicular to the surface of the conductor, and its end faces are parallel to the surface. Because the cylinder is infinitesimally small, the charge density σ is essentially constant over the surface enclosed, so the total charge inside the Gaussian cylinder is $\sigma \Delta A$. Now E is perpendicular to the surface of the conductor outside the conductor and vanishes within it, because otherwise, the charges would accelerate, and we would not be in equilibrium. Electric flux therefore crosses only the outer end face of the Gaussian surface and may be written as $E\Delta A$, since the cylinder is assumed to be small enough that E is approximately constant over that area. From Gauss' law,
Equation:

$$E\Delta A = \frac{\sigma \Delta A}{\varepsilon_0}.$$

Thus,
Equation:

$$E = \frac{\sigma}{\varepsilon_0}.$$

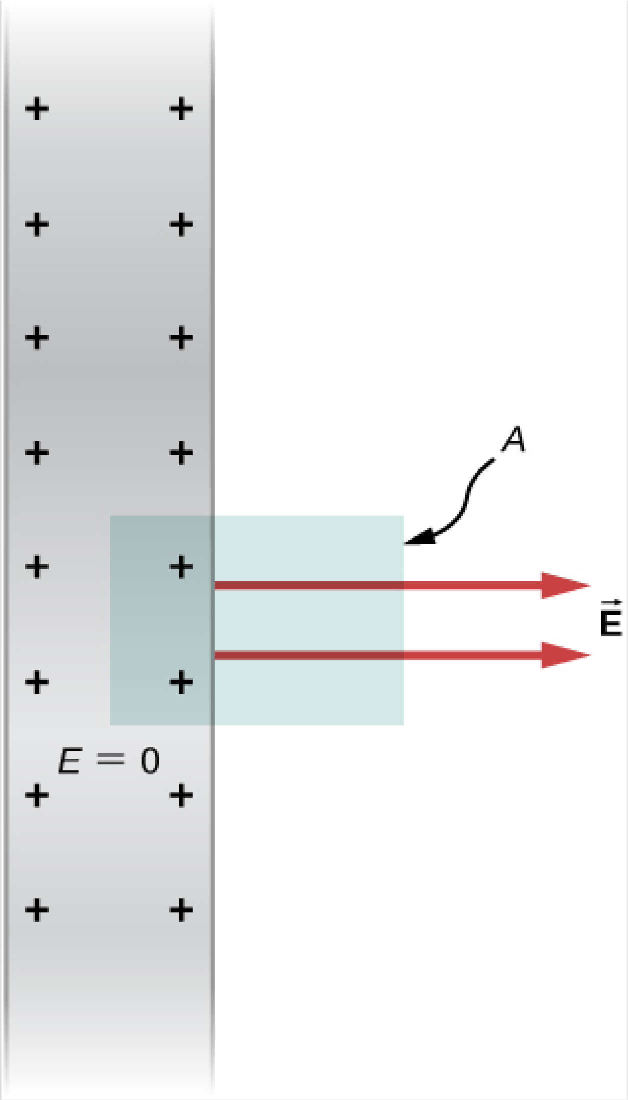


An infinitesimally small cylindrical Gaussian surface surrounds point P, which is on the surface of the conductor. The field $\vec{\mathbf{E}}$ is perpendicular to the surface of the conductor outside the conductor and vanishes within it.

Example:

Electric Field of a Conducting Plate

The infinite conducting plate in [link] has a uniform surface charge density σ. Use Gauss' law to find the electric field outside the plate. Compare this result with that previously calculated directly.



A side view of an infinite conducting plate and Gaussian cylinder with cross-sectional area *A*.

Strategy

For this case, we use a cylindrical Gaussian surface, a side view of which is shown.

Solution

The flux calculation is similar to that for an infinite sheet of charge from the previous chapter with one major exception: The left face of the Gaussian surface is inside the conductor where $\vec{\mathbf{E}} = \vec{\mathbf{0}}$, so the total flux through the Gaussian surface is *EA* rather than 2*EA*. Then from Gauss' law,

Equation:

$$EA = \frac{\sigma A}{\varepsilon_0}$$

and the electric field outside the plate is

Equation:

$$E = \frac{\sigma}{\varepsilon_0}.$$

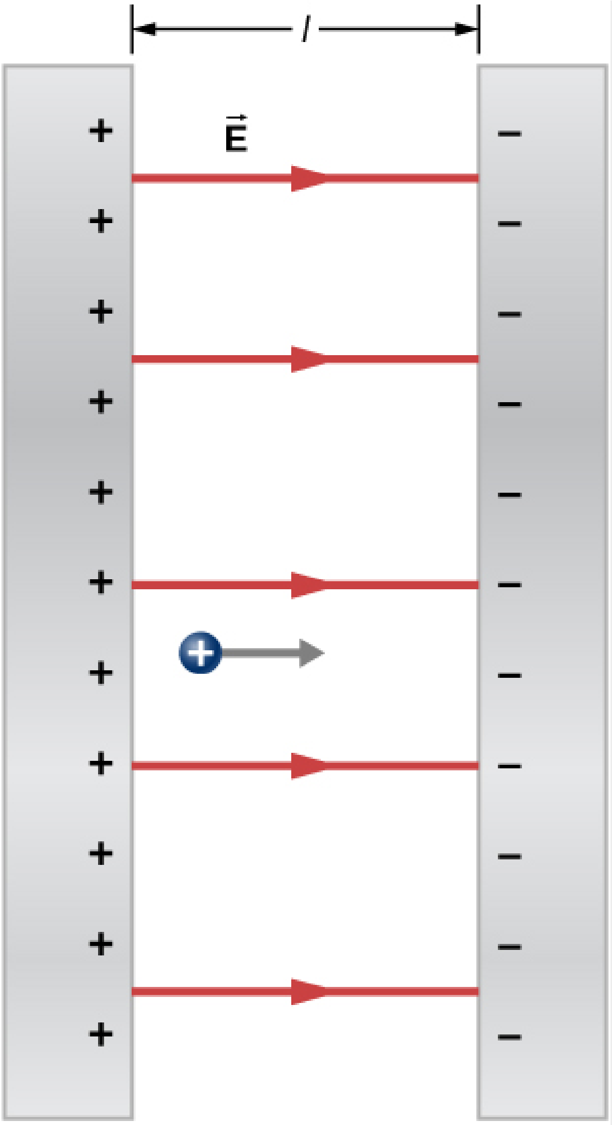
Significance

This result is in agreement with the result from the previous section, and consistent with the rule stated above.

Example:

Electric Field between Oppositely Charged Parallel Plates

Two large conducting plates carry equal and opposite charges, with a surface charge density σ of magnitude $6.81\ \times\ 10^{-7}\ \text{C/m}^2$, as shown in [link]. The separation between the plates is $l = 6.50\ \text{mm}$. What is the electric field between the plates?



The electric field between oppositely charged parallel plates. A test charge is released at the positive plate.

Strategy
Note that the electric field at the surface of one plate only depends on the charge on that plate. Thus, apply $E = \sigma/\varepsilon_0$ with the given values.
Solution

The electric field is directed from the positive to the negative plate, as shown in the figure, and its magnitude is given by

Equation:

$$E=\frac{\sigma}{\varepsilon_0}=\frac{6.81\ \times\ 10^{-7}\,\text{C/m}^2}{8.85\ \times\ 10^{-12}\,\text{C}^2/\text{N m}^2}=7.69\ \times\ 10^4\,\text{N/C}.$$

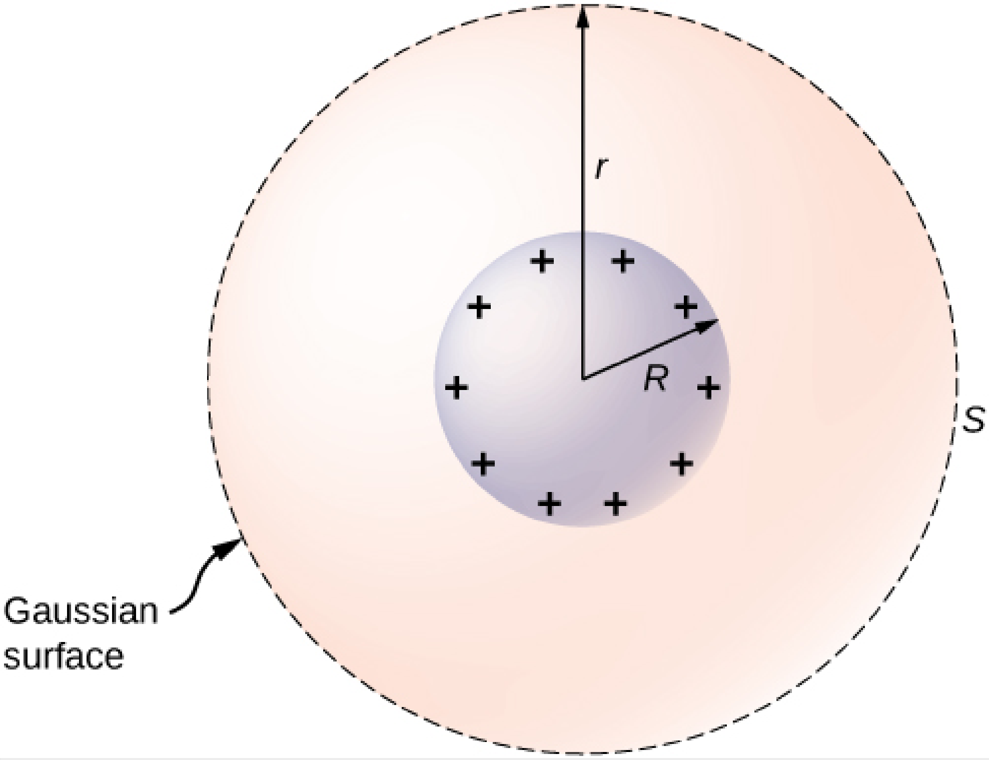
Significance

This formula is applicable to more than just a plate. Furthermore, two-plate systems will be important later.

Example:

A Conducting Sphere

The isolated conducting sphere ([link]) has a radius *R* and an excess charge *q*. What is the electric field both inside and outside the sphere?



An isolated conducting sphere.

Strategy
The sphere is isolated, so its surface change distribution and the electric field of that distribution are spherically symmetrical. We can therefore represent the field as $\overrightarrow{\mathbf{E}} = E(r)\hat{\mathbf{r}}$. To calculate $E(r)$, we apply Gauss's law over a closed spherical surface S of radius r that is concentric with the conducting sphere.

Solution
Since r is constant and $\hat{\mathbf{n}} = \hat{\mathbf{r}}$ on the sphere,

Equation:

$$\oint_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = E(r) \oint_S dA = E(r)\, 4\pi r^2.$$

For $r < R$, S is within the conductor, so $q_{\text{enc}} = 0$, and Gauss's law gives

Equation:

$$E\,(r) = 0,$$

as expected inside a conductor. If $r > R$, S encloses the conductor so $q_{\text{enc}} = q$. From Gauss's law,

Equation:

$$E(r)\, 4\pi r^2 = \frac{q}{\varepsilon_0}.$$

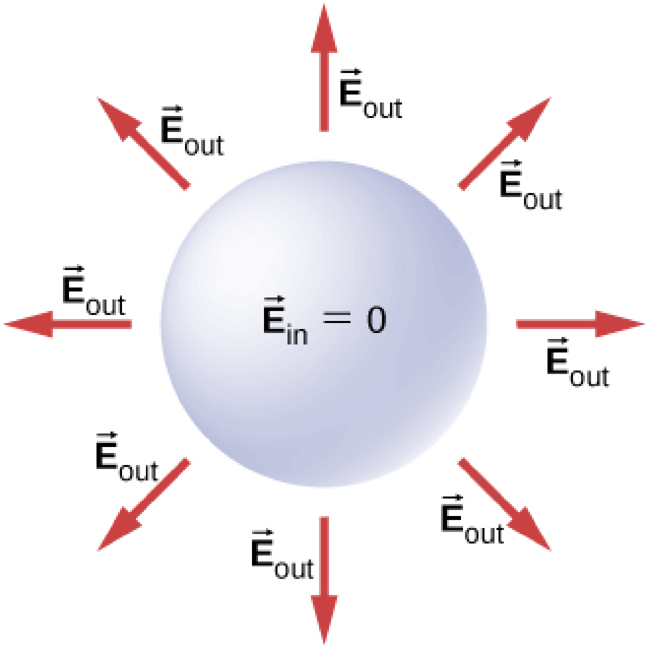
The electric field of the sphere may therefore be written as

Equation:

$$\begin{array}{lllr} \overrightarrow{\mathbf{E}} & = & \vec{\mathbf{0}} & (r < R), \\ \overrightarrow{\mathbf{E}} & = & \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}} & (r \geq R). \end{array}$$

Significance
Notice that in the region $r \geq R$, the electric field due to a charge q placed on an isolated conducting sphere of radius R is identical to the electric field of a point charge q located at the center of the sphere. The difference between the charged metal and a point charge occurs only at the space

points inside the conductor. For a point charge placed at the center of the sphere, the electric field is not zero at points of space occupied by the sphere, but a conductor with the same amount of charge has a zero electric field at those points ([link]). However, there is no distinction at the outside points in space where $r > R$, and we can replace the isolated charged spherical conductor by a point charge at its center with impunity.


$\vec{E}_{out}$
$\vec{E}_{out}$
$\vec{E}_{out}$
$\vec{E}_{out}$
$\vec{E}_{in} = 0$
$\vec{E}_{out}$
$\vec{E}_{out}$
$\vec{E}_{out}$
$\vec{E}_{out}$

Electric field of a positively charged metal sphere. The electric field inside is zero, and the electric field outside is same as the electric field of a point charge at the center, although the charge on the metal sphere is at the surface.

Note:
Exercise:

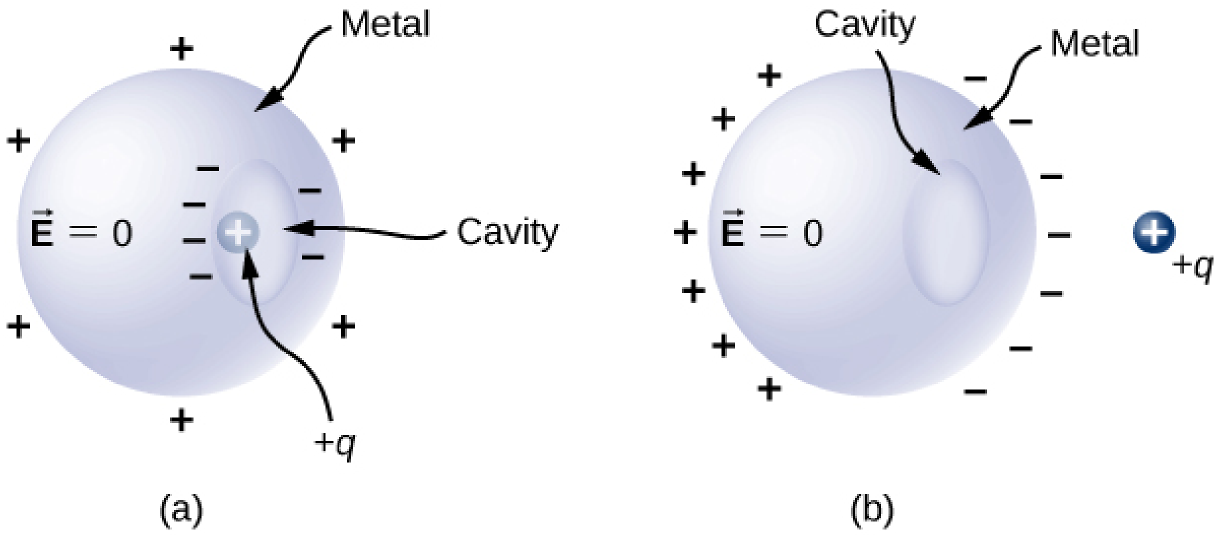
Problem:

Check Your Understanding How will the system above change if there are charged objects external to the sphere?

Solution:

If there are other charged objects around, then the charges on the surface of the sphere will not necessarily be spherically symmetrical; there will be more in certain direction than in other directions.

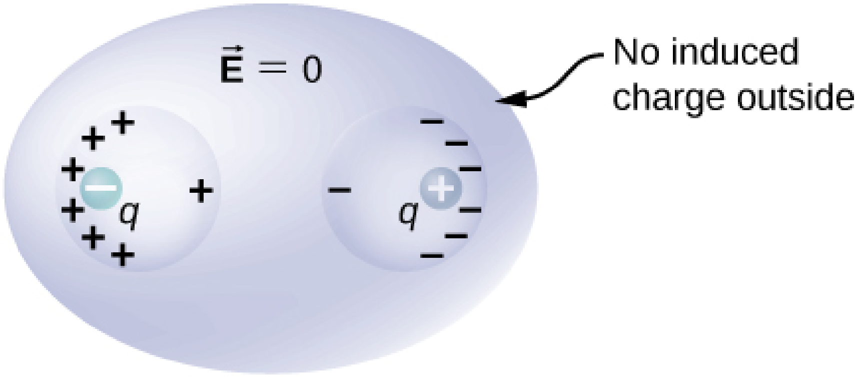
For a conductor with a cavity, if we put a charge $+q$ inside the cavity, then the charge separation takes place in the conductor, with $-q$ amount of charge on the inside surface and a $+q$ amount of charge at the outside surface ([link](a)). For the same conductor with a charge $+q$ outside it, there is no excess charge on the inside surface; both the positive and negative induced charges reside on the outside surface ([link](b)).



(a) A charge inside a cavity in a metal. The distribution of charges at the outer surface does not depend on how the charges are distributed at the inner surface, since the *E*-field inside the body of the metal is zero.

That magnitude of the charge on the outer surface does depend on the magnitude of the charge inside, however. (b) A charge outside a conductor containing an inner cavity. The cavity remains free of charge. The polarization of charges on the conductor happens at the surface.

If a conductor has two cavities, one of them having a charge $+q_a$ inside it and the other a charge $-q_b$, the polarization of the conductor results in $-q_a$ on the inside surface of the cavity a, $+q_b$ on the inside surface of the cavity b, and $q_a - q_b$ on the outside surface ([link]). The charges on the surfaces may not be uniformly spread out; their spread depends upon the geometry. The only rule obeyed is that when the equilibrium has been reached, the charge distribution in a conductor is such that the electric field by the charge distribution in the conductor cancels the electric field of the external charges at all space points inside the body of the conductor.



The charges induced by two equal and opposite charges in two separate cavities of a conductor. If the net charge on the cavity is nonzero, the external surface becomes charged to the amount of the net charge.

Summary

- The electric field inside a conductor vanishes.
- Any excess charge placed on a conductor resides entirely on the surface of the conductor.
- The electric field is perpendicular to the surface of a conductor everywhere on that surface.
- The magnitude of the electric field just above the surface of a conductor is given by $E = \frac{\sigma}{\varepsilon_0}$.

Key Equations

Definition of electric flux, for uniform electric field	$\Phi = \overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{A}} \rightarrow EA\cos\theta$
Electric flux through an open surface	$\Phi = \int_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \int_S \overrightarrow{\mathbf{E}} \cdot d\overrightarrow{\mathbf{A}}$
Electric flux through a closed surface	$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \oint_S \overrightarrow{\mathbf{E}} \cdot d\overrightarrow{\mathbf{A}}$
Gauss's law	$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \frac{q_{\text{enc}}}{\varepsilon_0}$
Gauss's Law for systems with symmetry	$\Phi = \oint_S \overrightarrow{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = E \oint_S dA = EA = \frac{q_{\text{enc}}}{\varepsilon_0}$
The magnitude of the electric field	$E = \frac{\sigma}{\varepsilon_0}$

just outside the surface of a conductor	

Conceptual Questions

Exercise:

Problem: Is the electric field inside a metal always zero?

Exercise:

Problem:

Under electrostatic conditions, the excess charge on a conductor resides on its surface. Does this mean that all the conduction electrons in a conductor are on the surface?

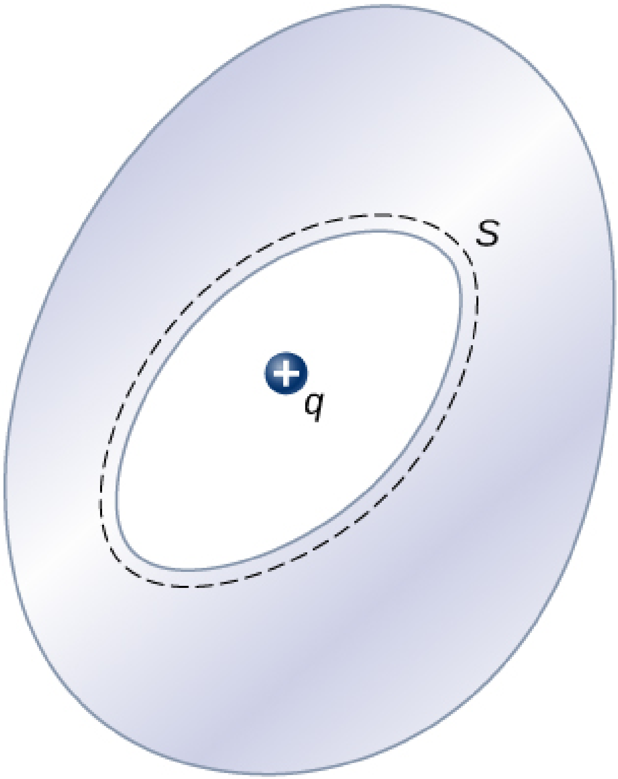
Solution:

No. If a metal was in a region of zero electric field, all the conduction electrons would be distributed uniformly throughout the metal.

Exercise:

Problem:

A charge q is placed in the cavity of a conductor as shown below. Will a charge outside the conductor experience an electric field due to the presence of q?



Exercise:

Problem:

The conductor in the preceding figure has an excess charge of $-5.0\ \mu\text{C}$. If a 2.0-μC point charge is placed in the cavity, what is the net charge on the surface of the cavity and on the outer surface of the conductor?

Solution:

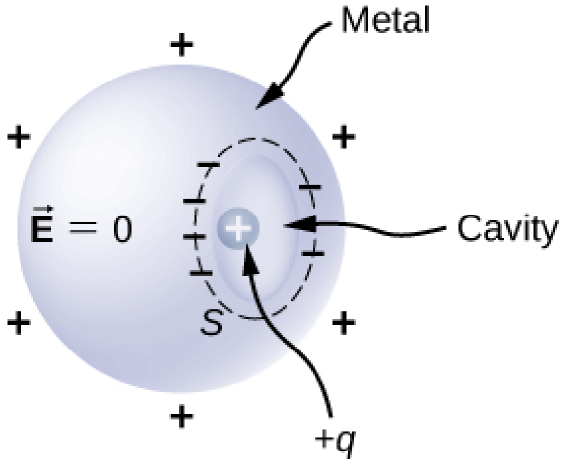
Since the electric field is zero inside a conductor, a charge of $-2.0\ \mu\text{C}$ is induced on the inside surface of the cavity. This will put a charge of $+2.0\ \mu C$ on the outside surface leaving a net charge of $-3.0\ \mu\text{C}$ on the surface.

Problems

Exercise:

Problem:

An uncharged conductor with an internal cavity is shown in the following figure. Use the closed surface *S* along with Gauss' law to show that when a charge *q* is placed in the cavity a total charge –*q* is induced on the inner surface of the conductor. What is the charge on the outer surface of the conductor?

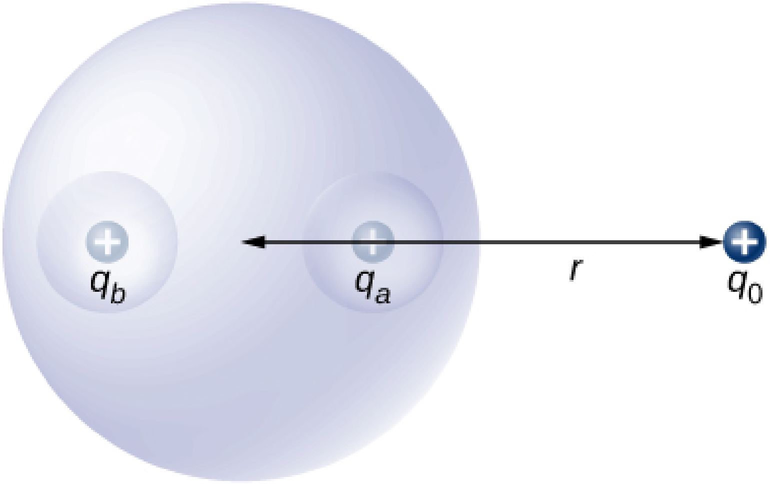


A charge inside a cavity of a metal. Charges at the outer surface do not depend on how the charges are distributed at the inner surface since E field inside the body of the metal is zero.

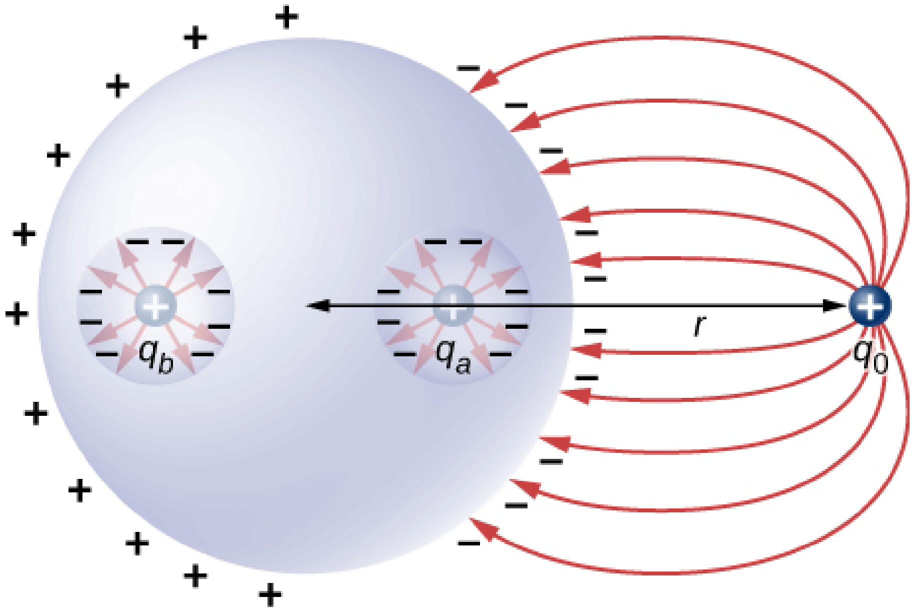
Exercise:

Problem:

An uncharged spherical conductor *S* of radius *R* has two spherical cavities A and B of radii *a* and *b*, respectively as shown below. Two point charges $+q_a$ and $+q_b$ are placed at the center of the two cavities by using non-conducting supports. In addition, a point charge $+q_0$ is placed outside at a distance *r* from the center of the sphere. (a) Draw approximate charge distributions in the metal although metal sphere has no net charge. (b) Draw electric field lines. Draw enough lines to represent all distinctly different places.



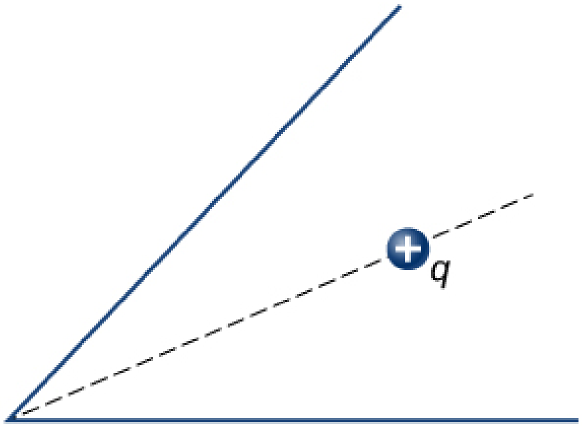
Solution:



Exercise:

Problem:

A positive point charge is placed at the angle bisector of two uncharged plane conductors that make an angle of 45°. See below. Draw the electric field lines.



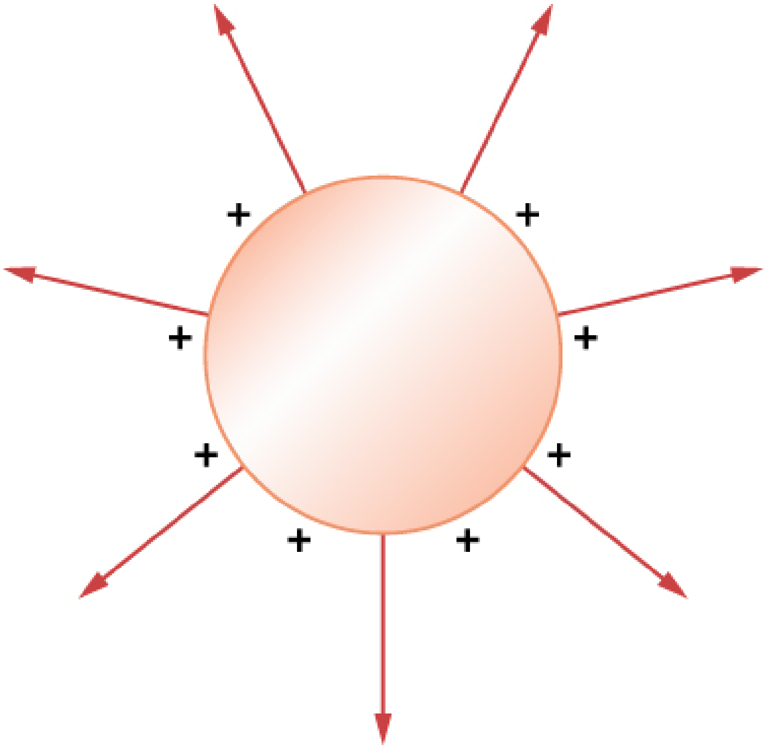
Exercise:

Problem:

A long cylinder of copper of radius 3 cm is charged so that it has a uniform charge per unit length on its surface of 3 C/m. (a) Find the electric field inside and outside the cylinder. (b) Draw electric field lines in a plane perpendicular to the rod.

Solution:

a. Outside: $E2\pi rl = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{3.0\,\text{C/m}}{2\pi\varepsilon_0 r}$; Inside $E_{\text{in}} = 0$; b.



Exercise:

Problem:

An aluminum spherical ball of radius 4 cm is charged with $5\ \mu\text{C}$ of charge. A copper spherical shell of inner radius 6 cm and outer radius 8 cm surrounds it. A total charge of $-8\ \mu\text{C}$ is put on the copper shell. (a) Find the electric field at all points in space, including points inside the aluminum and copper shell when copper shell and aluminum sphere are concentric. (b) Find the electric field at all points in space, including points inside the aluminum and copper shell when the centers of copper shell and aluminum sphere are 1 cm apart.

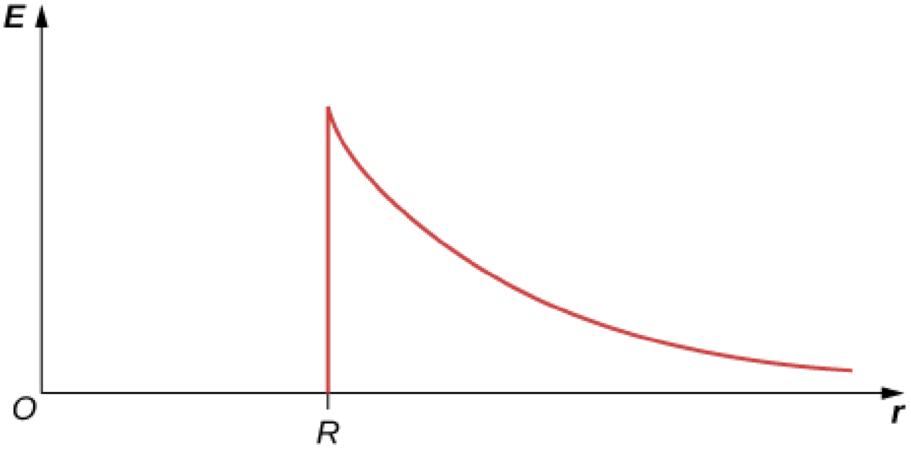
Exercise:

Problem:

A long cylinder of aluminum of radius *R* meters is charged so that it has a uniform charge per unit length on its surface of λ. (a) Find the electric field inside and outside the cylinder. (b) Plot electric field as a function of distance from the center of the rod.

Solution:

a. $E2\pi rl = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{\lambda}{2\pi\varepsilon_0 r}\ r \geq R$ *E* inside equals 0; b.



Exercise:

Problem:

At the surface of any conductor in electrostatic equilibrium, $E=\sigma/\varepsilon_0$. Show that this equation is consistent with the fact that $E=kq/r^2$ at the surface of a spherical conductor.

Exercise:

Problem:

Two parallel plates 10 cm on a side are given equal and opposite charges of magnitude $5.0\ \times\ 10^{-9}$ C. The plates are 1.5 mm apart. What is the electric field at the center of the region between the plates?

Solution:

$E=5.65\ \times\ 10^4\,\text{N/C}$

Exercise:

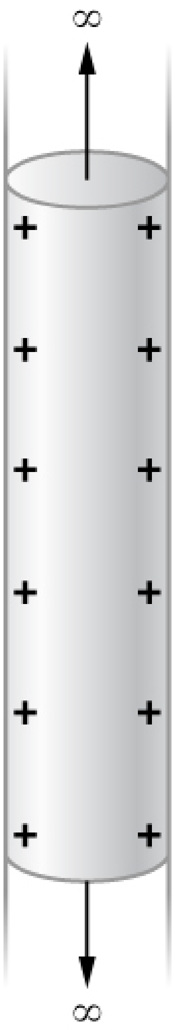
Problem:

Two parallel conducting plates, each of cross-sectional area $400\ \text{cm}^2$, are 2.0 cm apart and uncharged. If $1.0\ \times\ 10^{12}$ electrons are transferred from one plate to the other, what are (a) the charge density on each plate? (b) The electric field between the plates?

Exercise:

Problem:

The surface charge density on a long straight metallic pipe is σ. What is the electric field outside and inside the pipe? Assume the pipe has a diameter of $2a$.



Solution:

$\lambda = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{a\sigma}{\varepsilon_0 r}\, r \geq a, E = 0$ inside since q enclosed $= 0$

Exercise:

Problem:

A point charge $q = -5.0 \times 10^{-12}$ C is placed at the center of a spherical conducting shell of inner radius 3.5 cm and outer radius 4.0 cm. The electric field just above the surface of the conductor is directed radially outward and has magnitude 8.0 N/C. (a) What is the charge density on the inner surface of the shell? (b) What is the charge density on the outer surface of the shell? (c) What is the net charge on the conductor?

Exercise:

Problem:

A solid cylindrical conductor of radius *a* is surrounded by a concentric cylindrical shell of inner radius *b*. The solid cylinder and the shell carry charges +*Q* and –*Q*, respectively. Assuming that the length *L* of both conductors is much greater than *a* or *b*, determine the electric field as a function of *r*, the distance from the common central axis of the cylinders, for (a) $r < a$; (b) $a < r < b$; and (c)$r > b$.

Solution:

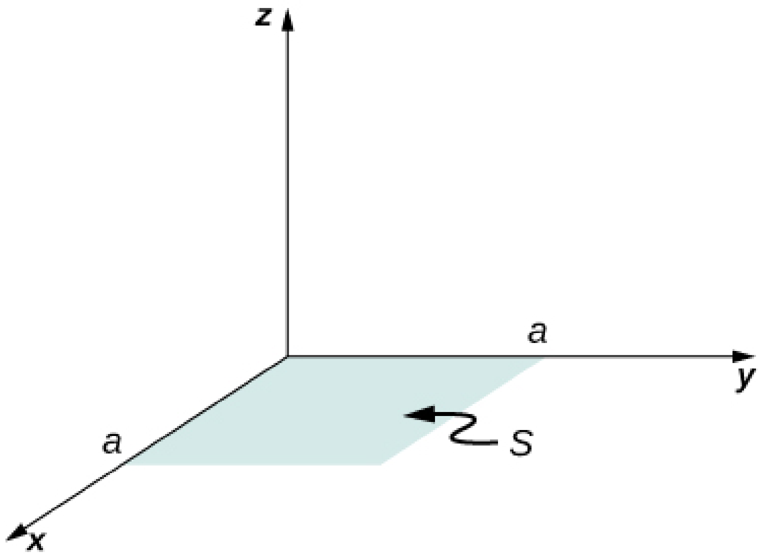
a. $E = 0$; b. $E2\pi rL = \frac{Q}{\varepsilon_0} \Rightarrow E = \frac{Q}{2\pi\varepsilon_0 rL}$; c. $E = 0$ since *r* would be either inside the second shell or if outside then *q* enclosed equals 0.

Additional Problems

Exercise:

Problem:

A vector field $\vec{\mathbf{E}}$ (not necessarily an electric field; note units) is given by $\vec{\mathbf{E}} = 3x^2\widehat{\mathbf{k}}$. Calculate $\int_S \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, da$, where *S* is the area shown below. Assume that $\widehat{\mathbf{n}} = \widehat{\mathbf{k}}$.



Exercise:

Problem: Repeat the preceding problem, with $\vec{\mathbf{E}} = 2x\hat{\mathbf{i}} + 3x^2\widehat{\mathbf{k}}$.

Solution:

$$\int \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = a^4$$

Exercise:

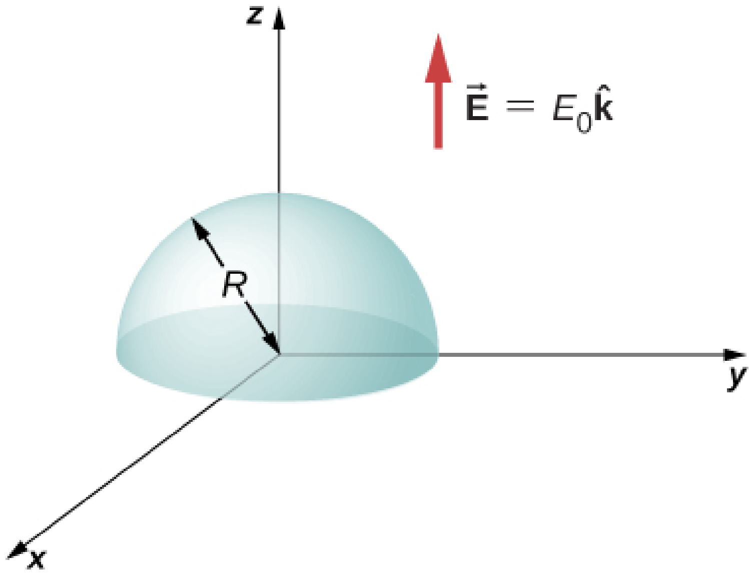
Problem:

A circular area *S* is concentric with the origin, has radius *a*, and lies in the *yz*-plane. Calculate $\int_S \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA$ for $\vec{\mathbf{E}} = 3z^2\hat{\mathbf{i}}$.

Exercise:

Problem:

(a) Calculate the electric flux through the open hemispherical surface due to the electric field $\vec{\mathbf{E}} = E_0\widehat{\mathbf{k}}$ (see below). (b) If the hemisphere is rotated by 90° around the *x*-axis, what is the flux through it?



Solution:

a. $\int \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, dA = E_0 r^2 \pi$; b. zero, since the flux through the upper half cancels the flux through the lower half of the sphere

Exercise:

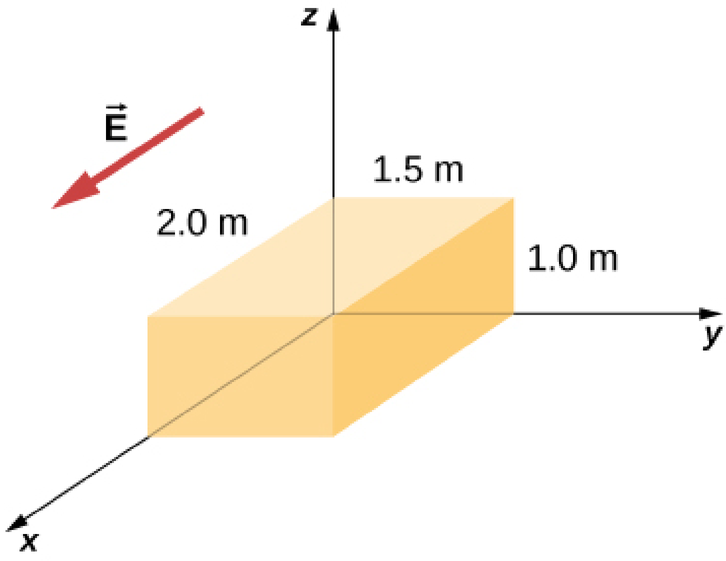
Problem:

Suppose that the electric field of an isolated point charge were proportional to $1/r^{2+\sigma}$ rather than $1/r^2$. Determine the flux that passes through the surface of a sphere of radius *R* centered at the charge. Would Gauss’s law remain valid?

Exercise:

Problem:

The electric field in a region is given by $\vec{\mathbf{E}} = a/(b + cx)\hat{\mathbf{i}}$, where $a = 200\,\text{N} \cdot \text{m/C}, b = 2.0\,\text{m}$, and $c = 2.0$. What is the net charge enclosed by the shaded volume shown below?



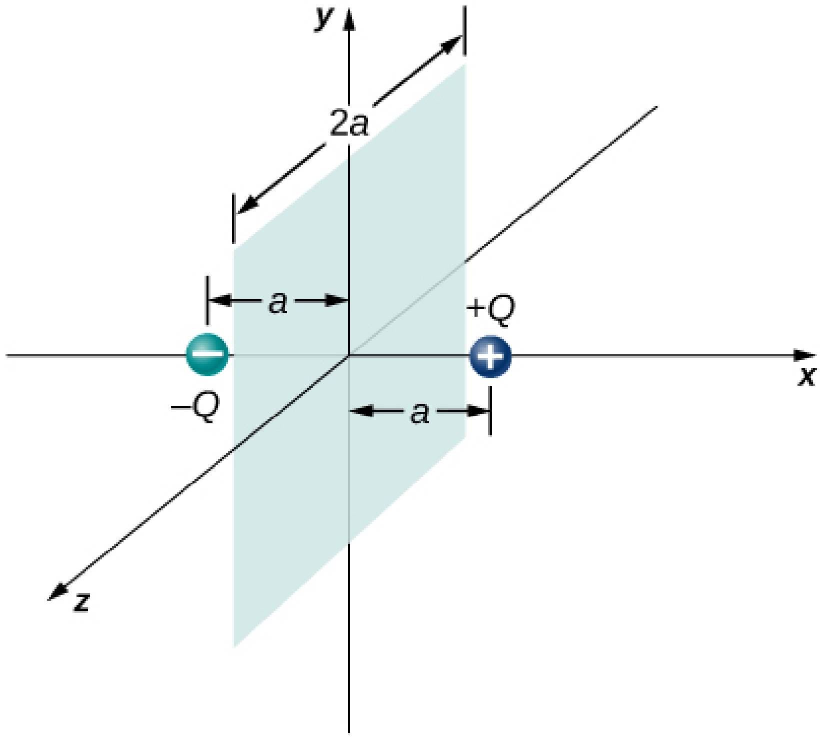
Solution:

$\Phi = \frac{q_{\text{enc}}}{\varepsilon_0}$; There are two contributions to the surface integral: one at the side of the rectangle at $x = 0$ and the other at the side at $x = 2.0\text{ m}$;
$-E(0)[1.5\text{ m}^2] + E(2.0\text{ m})[1.5\text{ m}^2] = \frac{q_{\text{enc}}}{\varepsilon_0} = -100\text{ Nm}^2/\text{C}$
where the minus sign indicates that at $x = 0$, the electric field is along positive *x* and the unit normal is along negative *x*. At $x = 2$, the unit normal and the electric field vector are in the same direction:
$q_{\text{enc}} = \varepsilon_0\,\Phi = -8.85\;\times\;10^{-10}\text{ C}.$

Exercise:

Problem:

Two equal and opposite charges of magnitude *Q* are located on the *x*-axis at the points +*a* and –*a*, as shown below. What is the net flux due to these charges through a square surface of side 2*a* that lies in the *yz*-plane and is centered at the origin? (*Hint:* Determine the flux due to each charge separately, then use the principle of superposition. You may be able to make a symmetry argument.)



Exercise:

Problem:

A fellow student calculated the flux through the square for the system in the preceding problem and got 0. What went wrong?

Solution:

didn’t keep consistent directions for the area vectors, or the electric fields

Exercise:

Problem:

A $10\,\text{cm}\ \times\ 10\,\text{cm}$ piece of aluminum foil of 0.1 mm thickness has a charge of $20\,\mu\text{C}$ that spreads on both wide side surfaces evenly. You may ignore the charges on the thin sides of the edges. (a) Find the charge density. (b) Find the electric field 1 cm from the center, assuming approximate planar symmetry.

Exercise:

Problem:

Two $10\text{ cm} \times 10\text{ cm}$ pieces of aluminum foil of thickness 0.1 mm face each other with a separation of 5 mm. One of the foils has a charge of $+30\ \mu\text{C}$ and the other has $-30\ \mu\text{C}$. (a) Find the charge density at all surfaces, i.e., on those facing each other and those facing away. (b) Find the electric field between the plates near the center assuming planar symmetry.

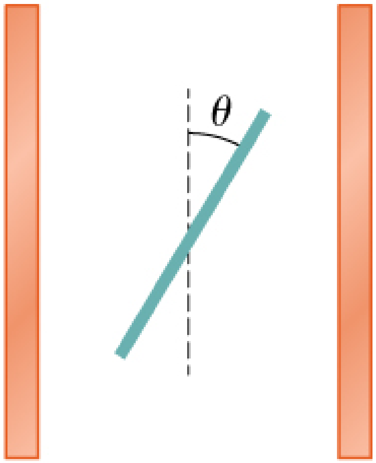
Solution:

a. $\sigma = 3.0 \times 10^{-3}\ \text{C/m}^2$, $+3 \times 10^{-3}\ \text{C/m}^2$ on one and $-3 \times 10^{-3}\ \text{C/m}^2$ on the other; b. $E = 3.39 \times 10^8\ \text{N/C}$

Exercise:

Problem:

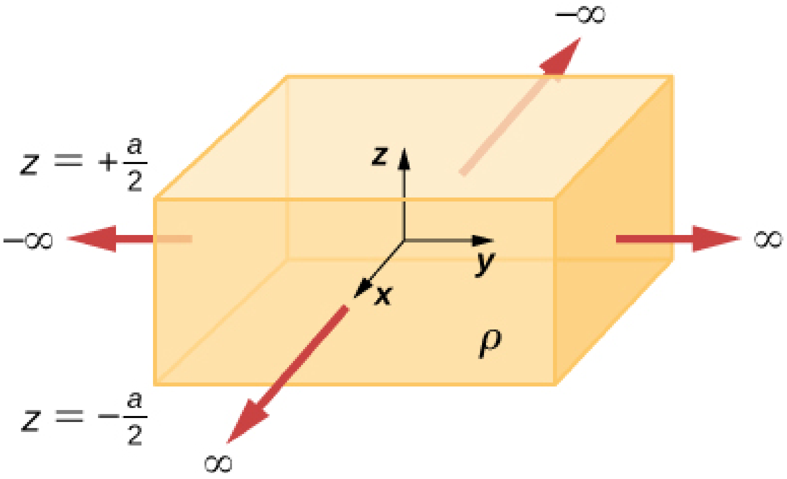
Two large copper plates facing each other have charge densities $\pm 4.0\ \text{C/m}^2$ on the surface facing the other plate, and zero in between the plates. Find the electric flux through a $3\text{ cm} \times 4\text{ cm}$ rectangular area between the plates, as shown below, for the following orientations of the area. (a) If the area is parallel to the plates, and (b) if the area is tilted $\theta = 30^\circ$ from the parallel direction. Note, this angle can also be $\theta = 180^\circ + 30^\circ$.



Exercise:

Problem:

The infinite slab between the planes defined by $z=-a/2$ and $z=a/2$ contains a uniform volume charge density ρ (see below). What is the electric field produced by this charge distribution, both inside and outside the distribution?



Solution:

Construct a Gaussian cylinder along the *z*-axis with cross-sectional area *A*.

$|z|\geq\frac{a}{2}\ q_{\text{enc}}=\rho Aa,\ \Phi=\frac{\rho Aa}{\varepsilon_0}\Rightarrow E=\frac{\rho a}{2\varepsilon_0},$

$|z|\leq\frac{a}{2}\ q_{\text{enc}}=\rho A2z,\ E(2A)=\frac{\rho A2z}{\varepsilon_0}\Rightarrow E=\frac{\rho z}{\varepsilon_0}$

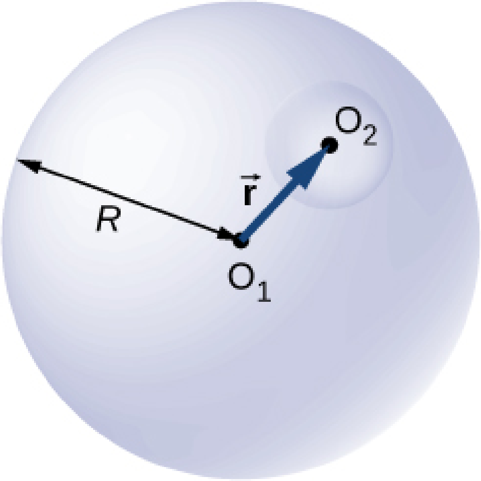
Exercise:

Problem:

A total charge Q is distributed uniformly throughout a spherical volume that is centered at O_1 and has a radius R. Without disturbing the charge remaining, charge is removed from the spherical volume that is centered at O_2 (see below). Show that the electric field everywhere in the empty region is given by

$$\vec{\mathbf{E}}=\frac{Q\vec{\mathbf{r}}}{4\pi\varepsilon_0 R^3},$$

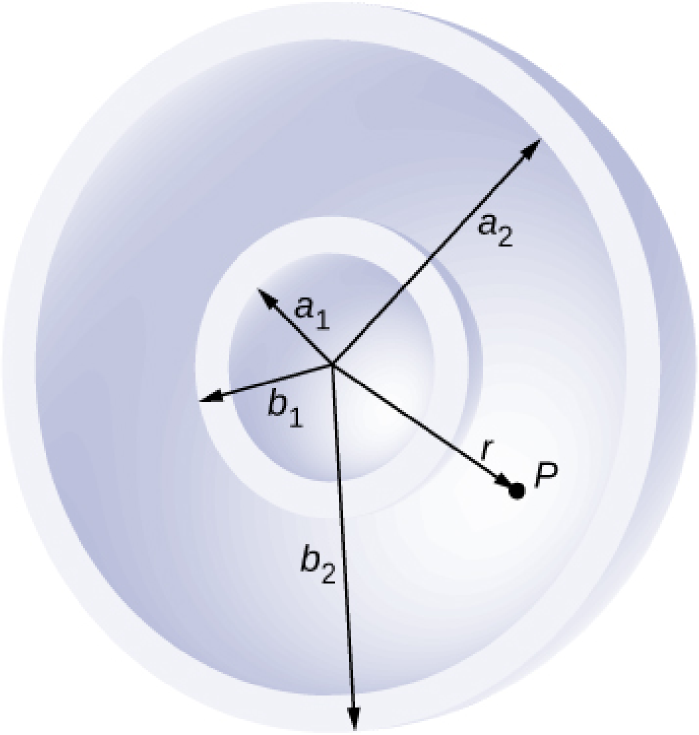
where $\vec{\mathbf{r}}$ is the displacement vector directed from O_1 to O_2.



Exercise:

Problem:

A non-conducting spherical shell of inner radius a_1 and outer radius b_1 is uniformly charged with charged density ρ_1 inside another non-conducting spherical shell of inner radius a_2 and outer radius b_2 that is also uniformly charged with charge density ρ_2. See below. Find the electric field at space point *P* at a distance *r* from the common center such that (a) $r > b_2$, (b) $a_2 < r < b_2$, (c) $b_1 < r < a_2$, (d) $a_1 < r < b_1$, and (e) $r < a_1$.



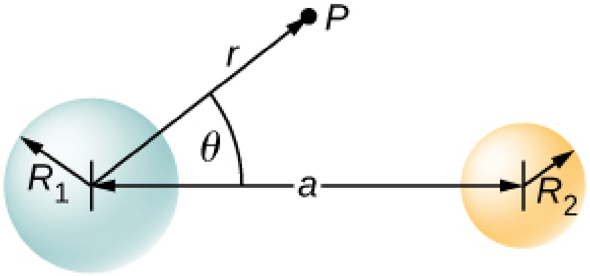
Solution:

a. $r > b_2$ $E4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3-a_1^3)+\rho_2(b_2^3-a_2^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3-a_1^3)+\rho_2(b_2^3-a_2^3)}{3\varepsilon_0 r^2}$;
b.
$a_2 < r < b_2 \quad E4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3-a_1^3)+\rho_2(r^3-a_2^3)]}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3-a_1^3)+\rho_2(r^3-a_2^3)}{3\varepsilon_0 r^2}$
;
c. $b_1 < r < a_2 \quad E4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(b_1^3-a_1^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3-a_1^3)}{3\varepsilon_0 r^2}$;
d. $a_1 < r < b_1 \quad E4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(r^3-a_1^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(r^3-a_1^3)}{3\varepsilon_0 r^2}$; e. 0

Exercise:

Problem:

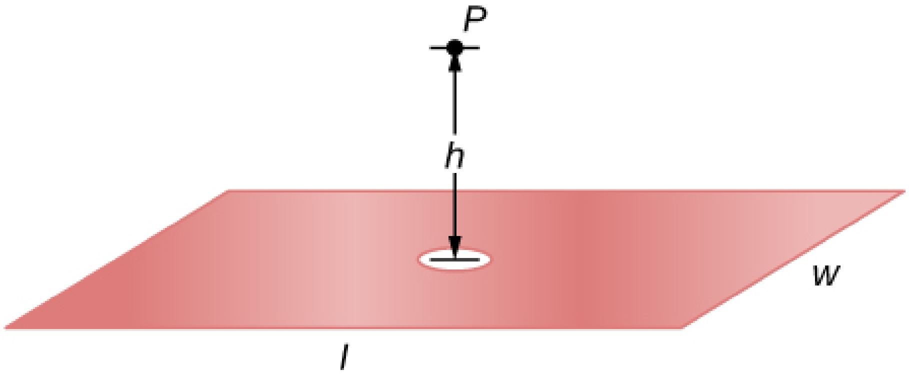
Two non-conducting spheres of radii R_1 and R_2 are uniformly charged with charge densities ρ_1 and ρ_2, respectively. They are separated at center-to-center distance *a* (see below). Find the electric field at point *P* located at a distance *r* from the center of sphere 1 and is in the direction θ from the line joining the two spheres assuming their charge densities are not affected by the presence of the other sphere. (*Hint:* Work one sphere at a time and use the superposition principle.)



Exercise:

Problem:

A disk of radius *R* is cut in a non-conducting large plate that is uniformly charged with charge density σ (coulomb per square meter). See below. Find the electric field at a height *h* above the center of the disk. $(h >> R, h << l$ or $w)$. (*Hint:* Fill the hole with $\pm\sigma$.)



Solution:

Electric field due to plate without hole: $E = \frac{\sigma}{2\varepsilon_0}$.

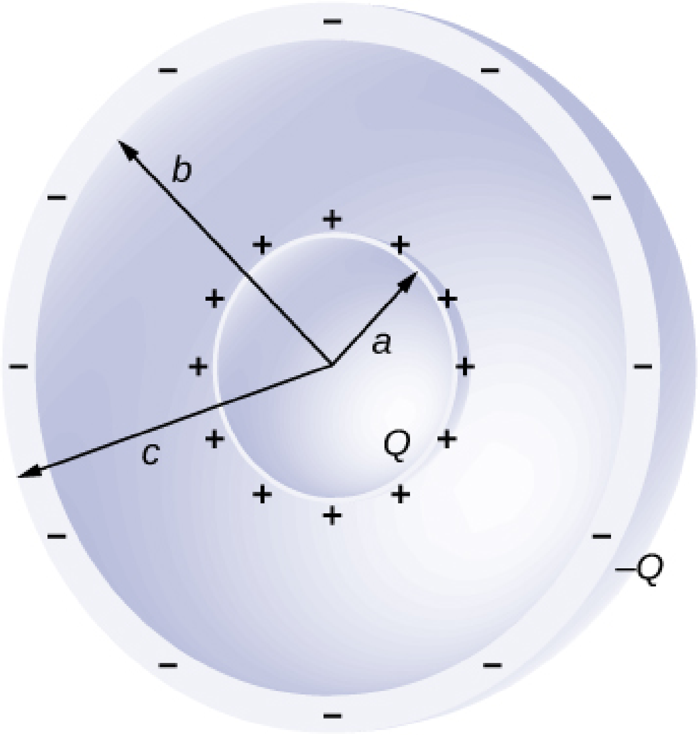
Electric field of just hole filled with $-\sigma$ $E = \frac{-\sigma}{2\varepsilon_0}\left(1 - \frac{z}{\sqrt{R^2+z^2}}\right)$.

Thus, $E_{\mathbf{net}} = \frac{\sigma}{2\varepsilon_0} \frac{h}{\sqrt{R^2+h^2}}$.

Exercise:

Problem:

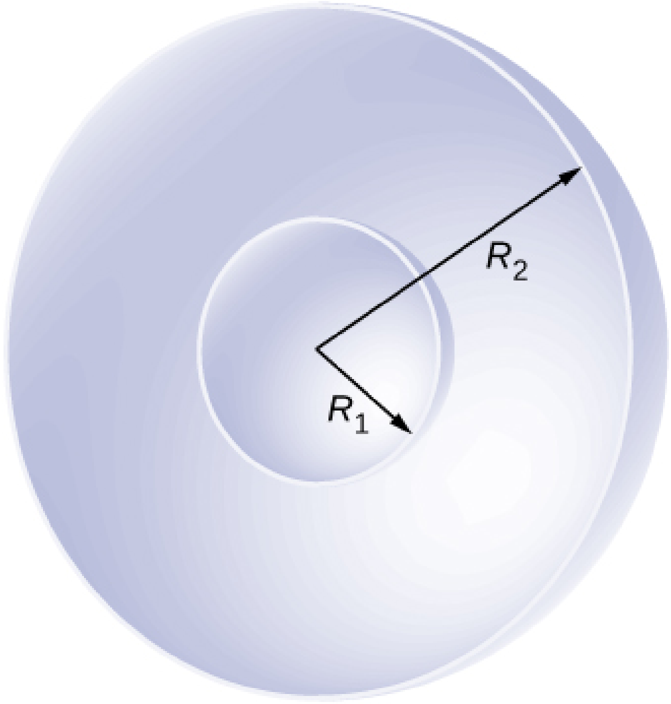
Concentric conducting spherical shells carry charges Q and –Q, respectively (see below). The inner shell has negligible thickness. Determine the electric field for (a) $r < a$; (b) $a < r < b$; (c) $b < r < c$; and (d) $r > c$.



Exercise:

Problem:

Shown below are two concentric conducting spherical shells of radii R_1 and R_2, each of finite thickness much less than either radius. The inner and outer shell carry net charges q_1 and q_2, respectively, where both q_1 and q_2 are positive. What is the electric field for (a) $r < R_1$; (b) $R_1 < r < R_2$; and (c) $r > R_2$? (d) What is the net charge on the inner surface of the inner shell, the outer surface of the inner shell, the inner surface of the outer shell, and the outer surface of the outer shell?



Solution:

a. $E = 0$; b. $E = \frac{q_1}{4\pi\varepsilon_0 r^2}$; c. $E = \frac{q_1+q_2}{4\pi\varepsilon_0 r^2}$; d. $0,\ q_1, -q_1,\ q_1 + q_2$

Exercise:

Problem:

A point charge of $q = 5.0 \times 10^{-8}\,\text{C}$ is placed at the center of an uncharged spherical conducting shell of inner radius 6.0 cm and outer radius 9.0 cm. Find the electric field at (a) $r = 4.0\,\text{cm}$, (b) $r = 8.0\,\text{cm}$, and (c) $r = 12.0\,\text{cm}$. (d) What are the charges induced on the inner and outer surfaces of the shell?

Challenge Problems

Exercise:

Problem:

The Hubble Space Telescope can measure the energy flux from distant objects such as supernovae and stars. Scientists then use this data to calculate the energy emitted by that object. Choose an interstellar object which scientists have observed the flux at the Hubble with (for example, Vega[footnote]), find the distance to that object and the size of Hubble's primary mirror, and calculate the total energy flux. (*Hint:* The Hubble intercepts only a small part of the total flux.)
http://adsabs.harvard.edu/abs/2004AJ....127.3508B

Solution:

Given the referenced link, using a distance to Vega of 237×10^{15} m[footnote] and a diameter of 2.4 m for the primary mirror,[footnote] we find that at a wavelength of 555.6 nm, Vega is emitting $2.44 \times 10^{24}\,\text{J/s}$ at that wavelength. Note that the flux through the mirror is essentially constant.
http://webviz.u-strasbg.fr/viz-bin/VizieR-5?-source=I/311&HIP=91262
http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19910003124.pdf

Exercise:

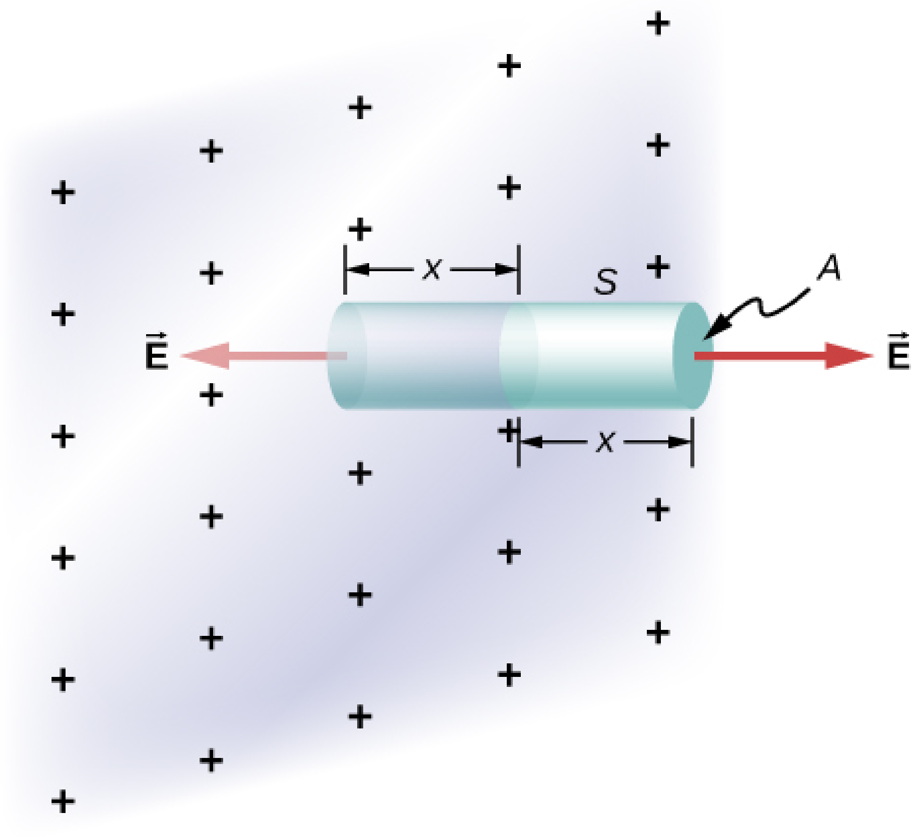
Problem:

Re-derive Gauss's law for the gravitational field, with $\vec{\mathbf{g}}$ directed positively outward.

Exercise:

Problem:

An infinite plate sheet of charge of surface charge density σ is shown below. What is the electric field at a distance *x* from the sheet? Compare the result of this calculation with that of worked out in the text.



Solution:

The symmetry of the system forces $\overrightarrow{\mathbf{E}}$ to be perpendicular to the sheet and constant over any plane parallel to the sheet. To calculate the

electric field, we choose the cylindrical Gaussian surface shown. The cross-section area and the height of the cylinder are A and $2x$, respectively, and the cylinder is positioned so that it is bisected by the plane sheet. Since E is perpendicular to each end and parallel to the side of the cylinder, we have EA as the flux through each end and there is no flux through the side. The charge enclosed by the cylinder is σA, so from Gauss's law, $2EA = \frac{\sigma A}{\varepsilon_0}$, and the electric field of an infinite sheet of charge is
$E = \frac{\sigma}{2\varepsilon_0}$, in agreement with the calculation of in the text.

Exercise:

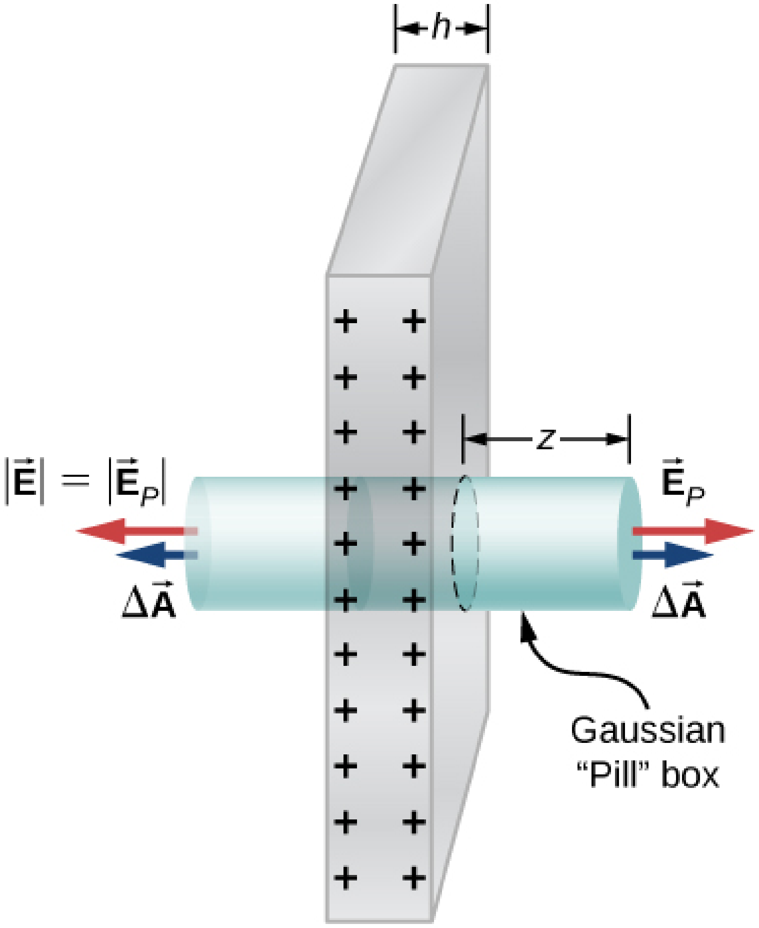
Problem:

A spherical rubber balloon carries a total charge Q distributed uniformly over its surface. At $t = 0$, the radius of the balloon is R. The balloon is then slowly inflated until its radius reaches $2R$ at the time t_0. Determine the electric field due to this charge as a function of time (a) at the surface of the balloon, (b) at the surface of radius R, and (c) at the surface of radius $2R$. Ignore any effect on the electric field due to the material of the balloon and assume that the radius increases uniformly with time.

Exercise:

Problem:

Find the electric field of a large conducting plate containing a net charge q. Let A be area of one side of the plate and h the thickness of the plate (see below). The charge on the metal plate will distribute mostly on the two planar sides and very little on the edges if the plate is thin.



h
$|\vec{\mathbf{E}}| = |\vec{\mathbf{E}}_P|$
z
$\vec{\mathbf{E}}_P$
$\Delta\vec{\mathbf{A}}$
$\Delta\vec{\mathbf{A}}$
Gaussian
"Pill" box

Solution:

There is $Q/2$ on each side of the plate since the net charge is Q: $\sigma = \frac{Q}{2A}$

,

$$\oint_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}}\, dA = \frac{2\sigma\Delta A}{\varepsilon_0} \Rightarrow E_P = \frac{\sigma}{\varepsilon_0} = \frac{Q}{\varepsilon_0 2A}$$

Glossary

free electrons
also called conduction electrons, these are the electrons in a conductor that are not bound to any particular atom, and hence are free to move around

Introduction class="introduction"

The energy released in a lightning strike is an excellent illustration of the vast quantities of energy that may be stored and released by an electric potential difference. In this chapter, we calculate just how much energy can be released in a lightning strike and how this varies with the height of the clouds from the ground. (credit: modification of work

by Anthony
Quintano)

In Electric Charges and Fields, we just scratched the surface (or at least rubbed it) of electrical phenomena. Two terms commonly used to describe electricity are its energy and *voltage*, which we show in this chapter is directly related to the potential energy in a system.

We know, for example, that great amounts of electrical energy can be stored in batteries, are transmitted cross-country via currents through power lines, and may jump from clouds to explode the sap of trees. In a similar manner, at the molecular level, ions cross cell membranes and transfer information.

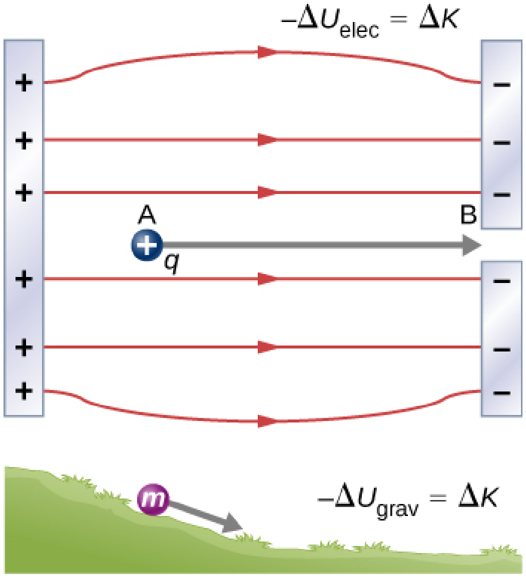
We also know about voltages associated with electricity. Batteries are typically a few volts, the outlets in your home frequently produce 120 volts, and power lines can be as high as hundreds of thousands of volts. But energy and voltage are not the same thing. A motorcycle battery, for example, is small and would not be very successful in replacing a much larger car battery, yet each has the same voltage. In this chapter, we examine the relationship between voltage and electrical energy, and begin to explore some of the many applications of electricity.

Electric Potential Energy

By the end of this section, you will be able to:

- Define the work done by an electric force
- Define electric potential energy
- Apply work and potential energy in systems with electric charges

When a free positive charge q is accelerated by an electric field, it is given kinetic energy ([link]). The process is analogous to an object being accelerated by a gravitational field, as if the charge were going down an electrical hill where its electric potential energy is converted into kinetic energy, although of course the sources of the forces are very different. Let us explore the work done on a charge q by the electric field in this process, so that we may develop a definition of electric potential energy.

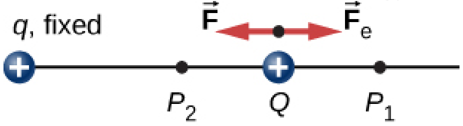


A charge accelerated by an electric field is analogous to a mass going down a hill. In both cases, potential energy decreases as kinetic energy increases, $-\Delta U = \Delta K$. Work is done by a force, but since this force is conservative, we can write $W = -\Delta U$.

The electrostatic or Coulomb force is conservative, which means that the work done on q is independent of the path taken, as we will demonstrate later. This is exactly analogous to the gravitational force. When a force is conservative, it is possible to define a potential energy associated with the force. It is usually easier to work with the potential energy (because it depends only on position) than to calculate the work directly.

To show this explicitly, consider an electric charge $+q$ fixed at the origin and move another charge $+Q$ toward q in such a manner that, at each instant, the applied force $\vec{\mathbf{F}}$ exactly balances the electric force

$\vec{\mathbf{F}}_e$ on Q ([link]). The work done by the applied force $\overrightarrow{\mathbf{F}}$ on the charge Q changes the potential energy of Q. We call this potential energy the **electrical potential energy** of Q.



Displacement of “test” charge Q in the presence of fixed “source” charge q.

The work W_{12} done by the applied force $\overrightarrow{\mathbf{F}}$ when the particle moves from P_1 to P_2 may be calculated by

Equation:

$$W_{12} = \int_{P_1}^{P_2} \overrightarrow{\mathbf{F}} \cdot d\vec{\mathbf{l}}.$$

Since the applied force $\overrightarrow{\mathbf{F}}$ balances the electric force $\vec{\mathbf{F}}_e$ on Q, the two forces have equal magnitude and opposite directions. Therefore, the applied force is

Equation:

$$\overrightarrow{\mathbf{F}} = -\mathbf{F}_e = -\frac{kqQ}{r^2}\hat{\mathbf{r}},$$

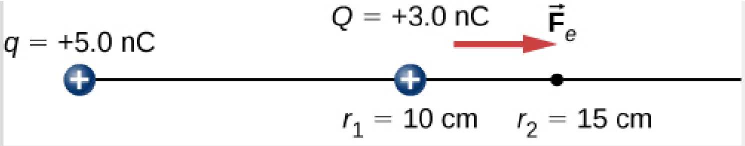
where we have defined positive to be pointing away from the origin and r is the distance from the origin. The directions of both the displacement and the applied force in the system in [link] are parallel, and thus the work done on the system is positive.

We use the letter U to denote electric potential energy, which has units of joules (J). When a conservative force does negative work, the system gains potential energy. When a conservative force does positive work, the system loses potential energy, $\Delta U = -W$. In the system in [link], the Coulomb force acts in the opposite direction to the displacement; therefore, the work is negative. However, we have increased the potential energy in the two-charge system.

Example:

Kinetic Energy of a Charged Particle

A $+3.0\text{-nC}$ charge Q is initially at rest a distance of 10 cm (r_1) from a $+5.0\text{-nC}$ charge q fixed at the origin ([link]). Naturally, the Coulomb force accelerates Q away from q, eventually reaching 15 cm (r_2).



The charge Q is repelled by q, thus having work done on it and gaining kinetic energy.

a. What is the work done by the electric field between r_1 and r_2?
b. How much kinetic energy does Q have at r_2?

Strategy
Calculate the work with the usual definition. Since Q started from rest, this is the same as the kinetic energy.

Solution
Integrating force over distance, we obtain

Equation:

$$\begin{aligned} W_{12} &= \int_{r_1}^{r_2} \vec{\mathbf{F}} \cdot d\vec{\mathbf{r}} = \int_{r_1}^{r_2} \frac{kqQ}{r^2}\,dr = \left[-\frac{kqQ}{r}\right]_{r_1}^{r_2} = kqQ\left[\frac{-1}{r_2} + \frac{1}{r_1}\right] \\ &= \left(8.99 \times 10^9\,\text{Nm}^2/\text{C}^2\right)\left(5.0 \times 10^{-9}\,\text{C}\right)\left(3.0 \times 10^{-9}\,\text{C}\right)\left[\frac{-1}{0.15\,\text{m}} + \frac{1}{0.10\,\text{m}}\right] \\ &= 4.5 \times 10^{-7}\,\text{J}. \end{aligned}$$

This is also the value of the kinetic energy at r_2.

Significance
Charge Q was initially at rest; the electric field of q did work on Q, so now Q has kinetic energy equal to the work done by the electric field.

Note:
Exercise:

Problem: Check Your Understanding If Q has a mass of 4.00 μg, what is the speed of Q at r_2?

Solution:

$$K = \frac{1}{2}mv^2,\ v = \sqrt{2\frac{K}{m}} = \sqrt{2\frac{4.5 \times 10^{-7}\,\text{J}}{4.00 \times 10^{-9}\,\text{kg}}} = 15\,\text{m/s}$$

In this example, the work W done to accelerate a positive charge from rest is positive and results from a loss in U, or a negative ΔU. A value for U can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Note:
Electric Potential Energy
Work W done to accelerate a positive charge from rest is positive and results from a loss in U, or a negative ΔU. Mathematically,
Equation:

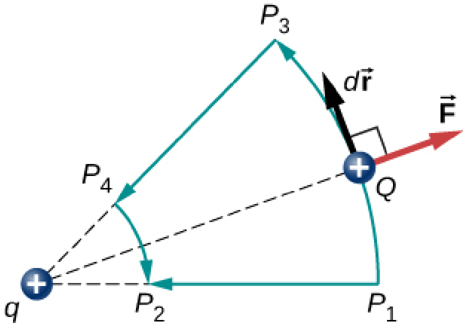
$$W = -\Delta U.$$

Gravitational potential energy and electric potential energy are quite analogous. Potential energy accounts for work done by a conservative force and gives added insight regarding energy and energy transformation without the necessity of dealing with the force directly. It is much more common, for example, to use the concept of electric potential energy than to deal with the Coulomb force directly in real-world applications.

In polar coordinates with q at the origin and Q located at r, the displacement element vector is $d\vec{\mathbf{l}} = \hat{\mathbf{r}}\, dr$ and thus the work becomes
Equation:

$$W_{12} = kqQ\int_{r_1}^{r_2} \frac{1}{r^2}\hat{\mathbf{r}} \cdot \hat{\mathbf{r}} dr = kqQ\frac{1}{r_2} - kqQ\frac{1}{r_1}.$$

Notice that this result only depends on the endpoints and is otherwise independent of the path taken. To explore this further, compare path P_1 to P_2 with path $P_1P_3P_4P_2$ in [link].



Two paths for displacement P_1 to P_2. The work on segments P_1P_3 and P_4P_2 are zero due to the electrical force being perpendicular to the displacement along these paths. Therefore, work on paths P_1P_2 and $P_1P_3P_4P_2$ are equal.

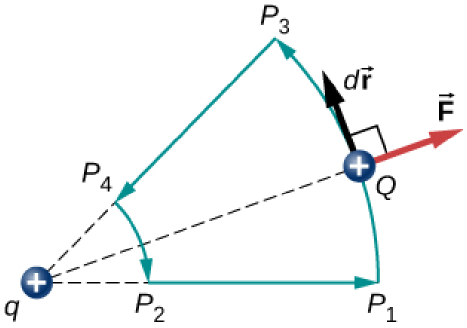
The segments P_1P_3 and P_4P_2 are arcs of circles centered at q. Since the force on Q points either toward or away from q, no work is done by a force balancing the electric force, because it is perpendicular to the displacement along these arcs. Therefore, the only work done is along segment P_3P_4, which is identical to P_1P_2.

One implication of this work calculation is that if we were to go around the path $P_1P_3P_4P_2P_1$, the net work would be zero ([link]). Recall that this is how we determine whether a force is conservative or not. Hence, because the electric force is related to the electric field by $\vec{\mathbf{F}} = q\vec{\mathbf{E}}$, the electric field is itself conservative. That is,

Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = 0.$$

Note that Q is a constant.



A closed path in an electric field. The net work around this path is zero.

Another implication is that we may define an electric potential energy. Recall that the work done by a conservative force is also expressed as the difference in the potential energy corresponding to that force. Therefore, the work W_{ref} to bring a charge from a reference point to a point of interest may be written as

Equation:

$$W_{\text{ref}} = \int_{r_{\text{ref}}}^{r} \vec{\mathbf{F}} \cdot d\vec{\mathbf{l}}$$

and, by [link], the difference in potential energy $(U_2 - U_1)$ of the test charge Q between the two points is

Equation:

$$\Delta U = -\int_{r_{\text{ref}}}^{r} \vec{\mathbf{F}} \cdot d\vec{\mathbf{l}}.$$

Therefore, we can write a general expression for the potential energy of two point charges (in spherical coordinates):

Equation:

$$\Delta U = -\int_{r_{\text{ref}}}^{r} \frac{kqQ}{r^2}\,dr = -\left[-\frac{kqQ}{r}\right]_{r_{\text{ref}}}^{r} = kqQ\left[\frac{1}{r} - \frac{1}{r_{\text{ref}}}\right].$$

We may take the second term to be an arbitrary constant reference level, which serves as the zero reference:

Equation:

$$U(r) = k\frac{qQ}{r} - U_{\text{ref}}.$$

A convenient choice of reference that relies on our common sense is that when the two charges are infinitely far apart, there is no interaction between them. (Recall the discussion of reference potential energy in Potential Energy and Conservation of Energy.) Taking the potential energy of this state to be zero removes the term U_{ref} from the equation (just like when we say the ground is zero potential energy in a gravitational potential energy problem), and the potential energy of Q when it is separated from q by a distance r assumes the form

Note:

Equation:

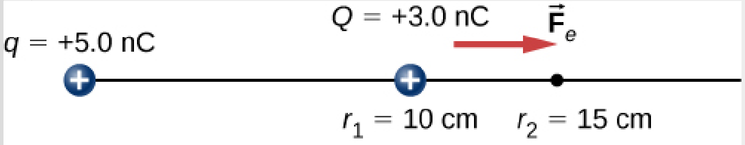
$$U(r) = k\frac{qQ}{r}\ (\text{zero reference at } r = \infty).$$

This formula is symmetrical with respect to q and Q, so it is best described as the potential energy of the two-charge system.

Example:

Potential Energy of a Charged Particle

A $+3.0\text{-nC}$ charge Q is initially at rest a distance of 10 cm (r_1) from a $+5.0\text{-nC}$ charge q fixed at the origin ([link]). Naturally, the Coulomb force accelerates Q away from q, eventually reaching 15 cm (r_2).



The charge Q is repelled by q, thus having work done on it and losing potential energy.

What is the change in the potential energy of the two-charge system from r_1 to r_2?

Strategy

Calculate the potential energy with the definition given above: $\Delta U_{12} = -\int_{r_1}^{r_2} \vec{\mathbf{F}} \cdot d\vec{\mathbf{r}}$. Since Q started from rest, this is the same as the kinetic energy.

Solution

We have

Equation:

$$\begin{aligned} \Delta U_{12} &= -\int_{r_1}^{r_2} \vec{\mathbf{F}} \cdot d\vec{\mathbf{r}} = -\int_{r_1}^{r_2} \frac{kqQ}{r^2} dr = -\left[-\frac{kqQ}{r}\right]_{r_1}^{r_2} = kqQ\left[\frac{1}{r_2} - \frac{1}{r_1}\right] \\ &= \left(8.99 \times 10^9\,\text{Nm}^2/\text{C}^2\right)\left(5.0 \times 10^{-9}\,\text{C}\right)\left(3.0 \times 10^{-9}\,\text{C}\right)\left[\frac{1}{0.15\,\text{m}} - \frac{1}{0.10\,\text{m}}\right] \\ &= -4.5 \times 10^{-7}\,\text{J}. \end{aligned}$$

Significance

The change in the potential energy is negative, as expected, and equal in magnitude to the change in kinetic energy in this system. Recall from [link] that the change in kinetic energy was positive.

Note:

Exercise:

Problem:

Check Your Understanding What is the potential energy of Q relative to the zero reference at infinity at r_2 in the above example?

Solution:

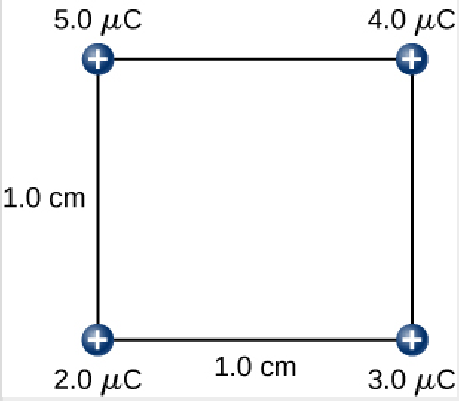
It has kinetic energy of 4.5×10^{-7} J at point r_2 and potential energy of 9.0×10^{-7} J, which means that as Q approaches infinity, its kinetic energy totals three times the kinetic energy at r_2, since all of the potential energy gets converted to kinetic.

Due to Coulomb's law, the forces due to multiple charges on a test charge Q superimpose; they may be calculated individually and then added. This implies that the work integrals and hence the resulting potential energies exhibit the same behavior. To demonstrate this, we consider an example of assembling a system of four charges.

Example:

Assembling Four Positive Charges

Find the amount of work an external agent must do in assembling four charges $+2.0\,\mu\text{C}$, $+3.0\,\mu\text{C}$, $+4.0\,\mu\text{C}$, and $+5.0\,\mu\text{C}$ at the vertices of a square of side 1.0 cm, starting each charge from infinity ([link]).



How much work is needed to assemble this charge configuration?

Strategy

We bring in the charges one at a time, giving them starting locations at infinity and calculating the work to bring them in from infinity to their final location. We do this in order of increasing charge.

Solution

Step 1. First bring the $+2.0$-μC charge to the origin. Since there are no other charges at a finite distance from this charge yet, no work is done in bringing it from infinity,

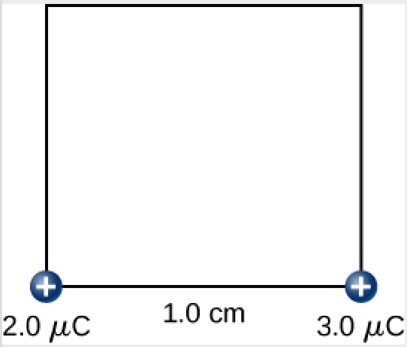
Equation:

$$W_1 = 0.$$

Step 2. While keeping the $+2.0$-μC charge fixed at the origin, bring the $+3.0$-μC charge to $(x, y, z) = (1.0\,\text{cm}, 0, 0)$ ([link]). Now, the applied force must do work against the force exerted by the $+2.0$-μC charge fixed at the origin. The work done equals the change in the potential energy of the $+3.0$-μC charge:

Equation:

$$W_2 = k\frac{q_1 q_2}{r_{12}} = \left(9.0 \times 10^9 \frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{\left(2.0 \times 10^{-6}\,\text{C}\right)\left(3.0 \times 10^{-6}\,\text{C}\right)}{1.0 \times 10^{-2}\,\text{m}} = 5.4\,\text{J}.$$



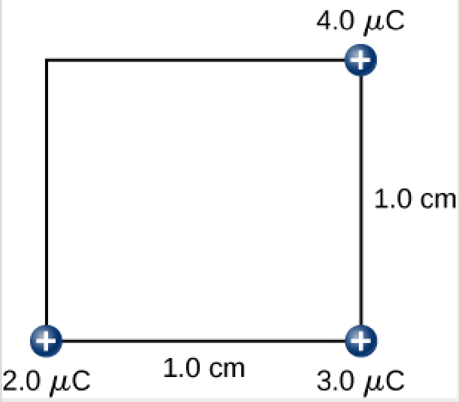
Step 2. Work W_2 to bring the $+3.0$-μC charge from

infinity.

Step 3. While keeping the charges of $+2.0\ \mu\text{C}$ and $+3.0\ \mu\text{C}$ fixed in their places, bring in the $+4.0\text{-}\mu\text{C}$ charge to $(x, y, z) = (1.0\ \text{cm}, 1.0\ \text{cm}, 0)$ ([link]). The work done in this step is

Equation:

$$\begin{aligned} W_3 &= k\frac{q_1q_3}{r_{13}} + k\frac{q_2q_3}{r_{23}} \\ &= \left(9.0\ \times\ 10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\left[\frac{\left(2.0\times10^{-6}\ \text{C}\right)\left(4.0\times10^{-6}\ \text{C}\right)}{\sqrt{2}\times10^{-2}\ \text{m}} + \frac{\left(3.0\times10^{-6}\ \text{C}\right)\left(4.0\times10^{-6}\ \text{C}\right)}{1.0\times10^{-2}\ \text{m}}\right] = 15.9\ \text{J}. \end{aligned}$$

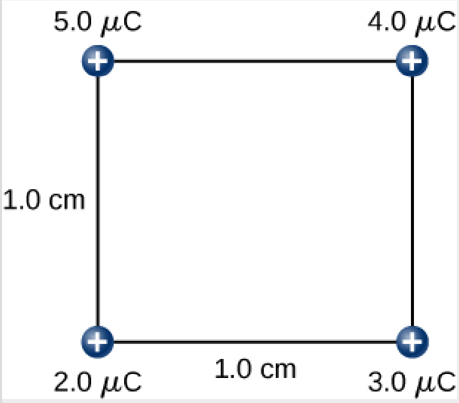


Step 3. The work W_3 to bring the $+4.0\text{-}\mu\text{C}$ charge from infinity.

Step 4. Finally, while keeping the first three charges in their places, bring the $+5.0\text{-}\mu\text{C}$ charge to $(x, y, z) = (0,\ 1.0\ \text{cm},\ 0)$ ([link]). The work done here is

Equation:

$$\begin{aligned} W_4 &= kq_4\left[\frac{q_1}{r_{14}} + \frac{q_2}{r_{24}} + \frac{q_3}{r_{34}}\right], \\ &= \left(9.0\ \times\ 10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\left(5.0\ \times\ 10^{-6}\ \text{C}\right)\left[\frac{\left(2.0\times10^{-6}\ \text{C}\right)}{1.0\times10^{-2}\ \text{m}} + \frac{\left(3.0\times10^{-6}\ \text{C}\right)}{\sqrt{2}\times10^{-2}\ \text{m}} + \frac{\left(4.0\times10^{-6}\ \text{C}\right)}{1.0\times10^{-2}\ \text{m}}\right] = 36.5\ \text{J}. \end{aligned}$$



Step 4. The work W_4 to bring the $+5.0$-μC charge from infinity.

Hence, the total work done by the applied force in assembling the four charges is equal to the sum of the work in bringing each charge from infinity to its final position:

Equation:

$$W_{\mathrm{T}} = W_1 + W_2 + W_3 + W_4 = 0 + 5.4\,\mathrm{J} + 15.9\,\mathrm{J} + 36.5\,\mathrm{J} = 57.8\,\mathrm{J}.$$

Significance

The work on each charge depends only on its pairwise interactions with the other charges. No more complicated interactions need to be considered; the work on the third charge only depends on its interaction with the first and second charges, the interaction between the first and second charge does not affect the third.

Note:

Exercise:

Problem:

Check Your Understanding Is the electrical potential energy of two point charges positive or negative if the charges are of the same sign? Opposite signs? How does this relate to the work necessary to bring the charges into proximity from infinity?

Solution:

positive, negative, and these quantities are the same as the work you would need to do to bring the charges in from infinity

Note that the electrical potential energy is positive if the two charges are of the same type, either positive or negative, and negative if the two charges are of opposite types. This makes sense if you think of the change in the potential energy ΔU as you bring the two charges closer or move them farther apart. Depending on the relative types of charges, you may have to work on the system or the system would do work on you, that is, your work is either positive or negative. If you have to do positive work on the system (actually push the charges closer), then the energy of the system should increase. If you bring two positive charges or two negative charges closer, you have to do positive work on the system, which raises their potential energy. Since potential energy is proportional to $1/r$, the potential energy goes up when r goes down between two positive or two negative charges.

On the other hand, if you bring a positive and a negative charge nearer, you have to do negative work on the system (the charges are pulling you), which means that you take energy away from the system. This reduces the potential energy. Since potential energy is negative in the case of a positive and a negative charge pair, the increase in $1/r$ makes the potential energy more negative, which is the same as a reduction in potential energy.

The result from [link] may be extended to systems with any arbitrary number of charges. In this case, it is most convenient to write the formula as

Note:
Equation:

$$W_{12\cdots N}=\frac{k}{2}\sum_{i}^{N}\sum_{j}^{N}\frac{q_iq_j}{r_{ij}}\text{ for }i\neq j.$$

The factor of 1/2 accounts for adding each pair of charges twice.

Summary

- The work done to move a charge from point *A* to *B* in an electric field is path independent, and the work around a closed path is zero. Therefore, the electric field and electric force are conservative.
- We can define an electric potential energy, which between point charges is $U(r)=k\frac{qQ}{r}$, with the zero reference taken to be at infinity.
- The superposition principle holds for electric potential energy; the potential energy of a system of multiple charges is the sum of the potential energies of the individual pairs.

Conceptual Questions

Exercise:

Problem:

Would electric potential energy be meaningful if the electric field were not conservative?

Solution:

No. We can only define potential energies for conservative fields.

Exercise:

Problem:

Why do we need to be careful about work done *on* the system versus work done *by* the system in calculations?

Exercise:

Problem:

Does the order in which we assemble a system of point charges affect the total work done?

Solution:

No, though certain orderings may be simpler to compute.

Problems

Exercise:

Problem:

Consider a charge $Q_1 (+5.0\ \mu\text{C})$ fixed at a site with another charge Q_2 (charge $+3.0\ \mu\text{C}$, mass $6.0\ \mu\text{g})$ moving in the neighboring space. (a) Evaluate the potential energy of Q_2 when it is 4.0 cm from Q_1. (b) If Q_2 starts from rest from a point 4.0 cm from Q_1, what will be its speed when it is 8.0 cm from Q_1? (*Note:* Q_1 is held fixed in its place.)

Solution:

a. $U = 3.4$ J;
b. $\frac{1}{2}mv^2 = Q_1 Q_2 \left(\frac{1}{r_i} - \frac{1}{r_f}\right) \rightarrow v = 2.4 \text{ x } 10^4\ \text{m/s}$

Exercise:

Problem:

Two charges $Q_1 (+2.00\ \mu\text{C})$ and $Q_2 (+2.00\ \mu\text{C})$ are placed symmetrically along the *x*-axis at $x = \pm 3.00\ \text{cm}$. Consider a charge Q_3 of charge $+4.00\ \mu\text{C}$ and mass 10.0 mg moving along the *y*-axis. If Q_3 starts from rest at $y = 2.00\ \text{cm}$, what is its speed when it reaches $y = 4.00\ \text{cm}$?

Exercise:

Problem:

To form a hydrogen atom, a proton is fixed at a point and an electron is brought from far away to a distance of $0.529\ \times\ 10^{-10}\ \text{m}$, the average distance between proton and electron in a hydrogen atom. How much work is done?

Solution:

$U = 4.36\ \times\ 10^{-18}$ J

Exercise:

Problem:

(a) What is the average power output of a heart defibrillator that dissipates 400 J of energy in 10.0 ms? (b) Considering the high-power output, why doesn’t the defibrillator produce serious burns?

Glossary

electric potential energy	potential energy stored in a system of charged objects due to the charges

Electric Potential and Potential Difference

By the end of this section, you will be able to:

- Define electric potential, voltage, and potential difference
- Define the electron-volt
- Calculate electric potential and potential difference from potential energy and electric field
- Describe systems in which the electron-volt is a useful unit
- Apply conservation of energy to electric systems

Recall that earlier we defined electric field to be a quantity independent of the test charge in a given system, which would nonetheless allow us to calculate the force that would result on an arbitrary test charge. (The default assumption in the absence of other information is that the test charge is positive.) We briefly defined a field for gravity, but gravity is always attractive, whereas the electric force can be either attractive or repulsive. Therefore, although potential energy is perfectly adequate in a gravitational system, it is convenient to define a quantity that allows us to calculate the work on a charge independent of the magnitude of the charge. Calculating the work directly may be difficult, since $W = \overrightarrow{\mathbf{F}} \cdot \vec{\mathbf{d}}$ and the direction and magnitude of $\overrightarrow{\mathbf{F}}$ can be complex for multiple charges, for odd-shaped objects, and along arbitrary paths. But we do know that because $\overrightarrow{\mathbf{F}} = q\overrightarrow{\mathbf{E}}$, the work, and hence ΔU, is proportional to the test charge q. To have a physical quantity that is independent of test charge, we define **electric potential** V (or simply potential, since electric is understood) to be the potential energy per unit charge:

Note:

Electric Potential

The electric potential energy per unit charge is

Equation:

$$V = \frac{U}{q}.$$

Since U is proportional to q, the dependence on q cancels. Thus, V does not depend on q. The change in potential energy ΔU is crucial, so we are concerned with the difference in potential or potential difference ΔV between two points, where

Equation:

$$\Delta V = V_B - V_A = \frac{\Delta U}{q}.$$

Note:
Electric Potential Difference
The **electric potential difference** between points *A* and *B*, $V_B - V_A$, is defined to be the change in potential energy of a charge *q* moved from *A* to *B*, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.
Equation:

$$1\ \mathrm{V} = 1\ \mathrm{J/C}$$

The familiar term **voltage** is the common name for electric potential difference. Keep in mind that whenever a voltage is quoted, it is understood to be the potential difference between two points. For example, every battery has two terminals, and its voltage is the potential difference between them. More fundamentally, the point you choose to be zero volts is arbitrary. This is analogous to the fact that gravitational potential energy has an arbitrary zero, such as sea level or perhaps a lecture hall floor. It is worthwhile to emphasize the distinction between potential difference and electrical potential energy.

Note:
Potential Difference and Electrical Potential Energy
The relationship between potential difference (or voltage) and electrical potential energy is given by
Equation:

$$\Delta V = \frac{\Delta U}{q} \text{ or } \Delta U = q\Delta V.$$

Voltage is not the same as energy. Voltage is the energy per unit charge. Thus, a motorcycle battery and a car battery can both have the same voltage (more precisely, the same potential difference between battery terminals), yet one stores much more energy than the other because $\Delta U = q\Delta V$. The car battery can move more charge than the motorcycle battery, although both are 12-V batteries.

Example:

Calculating Energy

You have a 12.0-V motorcycle battery that can move 5000 C of charge, and a 12.0-V car battery that can move 60,000 C of charge. How much energy does each deliver? (Assume that the numerical value of each charge is accurate to three significant figures.)

Strategy

To say we have a 12.0-V battery means that its terminals have a 12.0-V potential difference. When such a battery moves charge, it puts the charge through a potential difference of 12.0 V, and the charge is given a change in potential energy equal to $\Delta U = q\Delta V$. To find the energy output, we multiply the charge moved by the potential difference.

Solution

For the motorcycle battery, $q = 5000 \text{ C}$ and $\Delta V = 12.0 \text{ V}$. The total energy delivered by the motorcycle battery is

Equation:

$$\Delta U_{\text{cycle}} = (5000 \text{ C})(12.0 \text{ V}) = (5000 \text{ C})(12.0 \text{ J/C}) = 6.00 \times 10^4 \text{ J}.$$

Similarly, for the car battery, $q = 60{,}000 \text{ C}$ and

Equation:

$$\Delta U_{\text{car}} = (60{,}000 \text{ C})(12.0 \text{ V}) = 7.20 \times 10^5 \text{ J}.$$

Significance

Voltage and energy are related, but they are not the same thing. The voltages of the batteries are identical, but the energy supplied by each is quite different. A car battery has a much larger engine to start than a motorcycle. Note also that as a battery is discharged, some of its energy is used internally and its terminal voltage drops, such as when headlights dim because of a depleted car battery. The energy supplied by the battery is still calculated as in this example, but not all of the energy is available for external use.

Note:

Exercise:

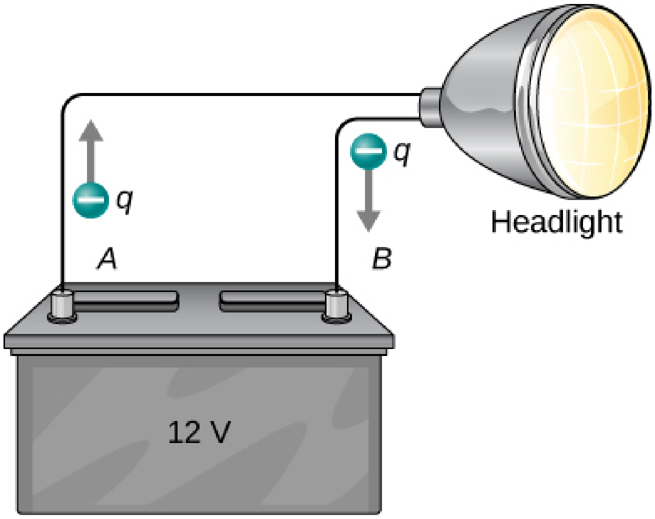
Problem:

Check Your Understanding How much energy does a 1.5-V AAA battery have that can move 100 C?

Solution:

$\Delta U = q\Delta V = (100 \text{ C})(1.5 \text{ V}) = 150 \text{ J}$

Note that the energies calculated in the previous example are absolute values. The change in potential energy for the battery is negative, since it loses energy. These batteries, like many electrical systems, actually move negative charge—electrons in particular. The batteries repel electrons from their negative terminals (A) through whatever circuitry is involved and attract them to their positive terminals (B), as shown in [link]. The change in potential is $\Delta V = V_B - V_A = +12\text{ V}$ and the charge q is negative, so that $\Delta U = q\Delta V$ is negative, meaning the potential energy of the battery has decreased when q has moved from A to B.



A battery moves negative charge from its negative terminal through a headlight to its positive terminal. Appropriate combinations of chemicals in the battery separate charges so that the negative terminal has an excess of negative charge, which is repelled by it and attracted to the excess positive charge on the other terminal. In terms of potential, the positive terminal is at a higher voltage than the negative terminal. Inside the battery, both positive and negative charges move.

Example:

How Many Electrons Move through a Headlight Each Second?

When a 12.0-V car battery powers a single 30.0-W headlight, how many electrons pass through it each second?

Strategy

To find the number of electrons, we must first find the charge that moves in 1.00 s. The charge moved is related to voltage and energy through the equations $\Delta U = q\Delta V$. A 30.0-W lamp uses 30.0 joules per second. Since the battery loses energy, we have $\Delta U = -30\text{ J}$ and, since the electrons are going from the negative terminal to the positive, we see that $\Delta V = +12.0\text{ V}$.

Solution

To find the charge q moved, we solve the equation $\Delta U = q\Delta V$:

Equation:

$$q = \frac{\Delta U}{\Delta V}.$$

Entering the values for ΔU and ΔV, we get

Equation:

$$q = \frac{-30.0\text{ J}}{+12.0\text{ V}} = \frac{-30.0\text{ J}}{+12.0\text{ J/C}} = -2.50\text{ C}.$$

The number of electrons n_e is the total charge divided by the charge per electron. That is,

Equation:

$$n_e = \frac{-2.50\text{ C}}{-1.60 \times 10^{-19}\text{ C/e}^-} = 1.56 \times 10^{19}\text{ electrons.}$$

Significance

This is a very large number. It is no wonder that we do not ordinarily observe individual electrons with so many being present in ordinary systems. In fact, electricity had been in use for many decades before it was determined that the moving charges in many circumstances were negative. Positive charge moving in the opposite direction of negative charge often produces identical effects; this makes it difficult to determine which is moving or whether both are moving.

Note:

Exercise:

Problem:

Check Your Understanding How many electrons would go through a 24.0-W lamp?

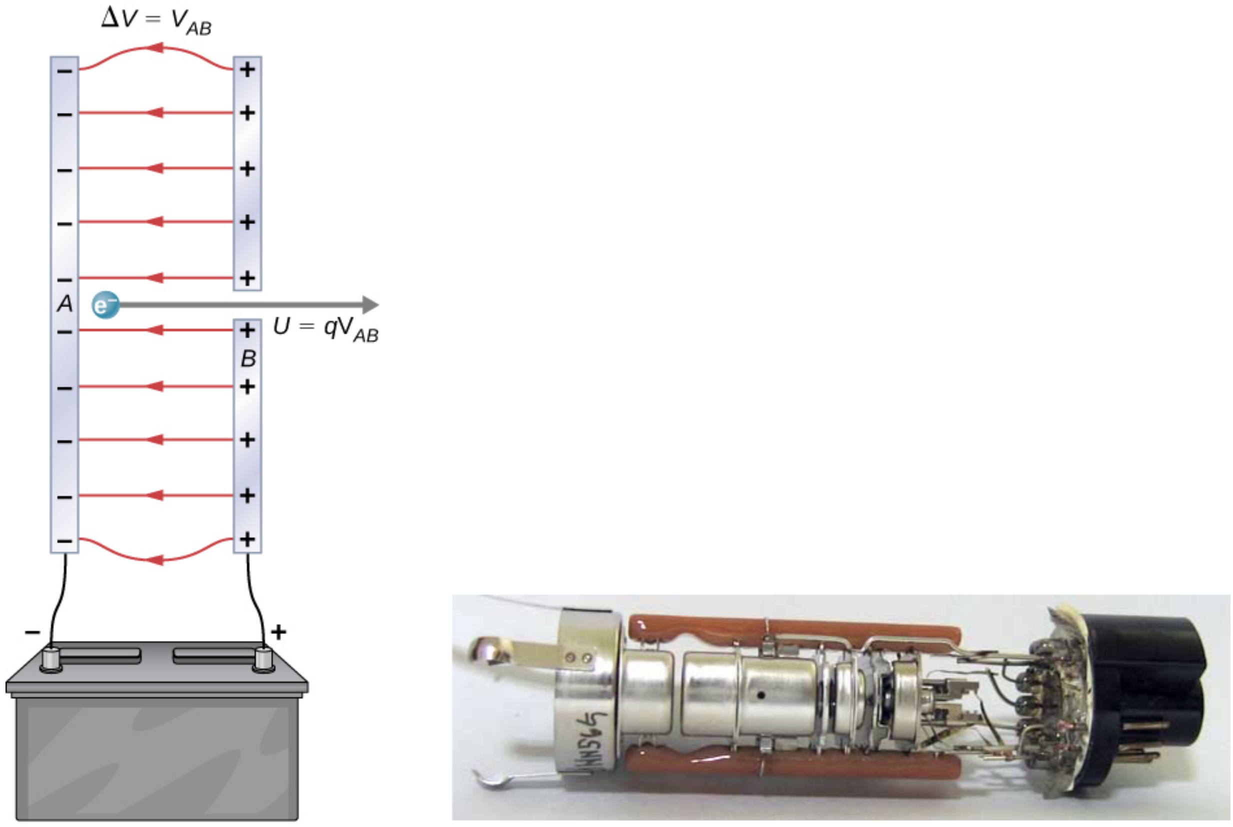
Solution:

–2.00 C, $n_e = 1.25 \times 10^{19}$ electrons

The Electron-Volt

The energy per electron is very small in macroscopic situations like that in the previous example—a tiny fraction of a joule. But on a submicroscopic scale, such energy per particle (electron, proton, or ion) can be of great importance. For example, even a tiny fraction of a joule can be great enough for these particles to destroy organic molecules and harm living tissue. The particle may do its damage by direct collision, or it may create harmful X-rays, which can also inflict damage. It is useful to have an energy unit related to submicroscopic effects.

[link] shows a situation related to the definition of such an energy unit. An electron is accelerated between two charged metal plates, as it might be in an old-model television tube or oscilloscope. The electron gains kinetic energy that is later converted into another form—light in the television tube, for example. (Note that in terms of energy, “downhill” for the electron is “uphill” for a positive charge.) Since energy is related to voltage by $\Delta U = q\Delta V$, we can think of the joule as a coulomb-volt.



(a)

(b)

A typical electron gun accelerates electrons using a potential difference between two separated metal plates. By conservation of energy, the kinetic energy has to equal the change in potential energy, so $KE = qV$. The energy of the electron in electron-volts is numerically the same as the voltage between the plates. For example, a 5000-V potential difference produces 5000-eV electrons. The conceptual construct, namely two parallel plates with a hole in one, is shown in (a), while a real electron gun is shown in (b).

Note:
Electron-Volt
On the submicroscopic scale, it is more convenient to define an energy unit called the **electron-volt** (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,
Equation:

$$1\text{ eV} = (1.60\ \times\ 10^{-19}\text{ C})(1\text{ V}) = (1.60\ \times\ 10^{-19}\text{ C})(1\text{ J/C}) = 1.60\ \times\ 10^{-19}\text{ J.}$$

An electron accelerated through a potential difference of 1 V is given an energy of 1 eV. It follows that an electron accelerated through 50 V gains 50 eV. A potential difference of 100,000 V (100 kV) gives an electron an energy of 100,000 eV (100 keV), and so on. Similarly, an ion with a double positive charge accelerated through 100 V gains 200 eV of energy. These simple relationships between accelerating voltage and particle charges make the electron-volt a simple and convenient energy unit in such circumstances.

The electron-volt is commonly employed in submicroscopic processes—chemical valence energies and molecular and nuclear binding energies are among the quantities often expressed in electron-volts. For example, about 5 eV of energy is required to break up certain organic molecules. If a proton is accelerated from rest through a potential difference of 30 kV, it acquires an energy of 30 keV (30,000 eV) and can break up as many as 6000 of these molecules $(30{,}000\text{ eV} \div 5\text{ eV per molecule} = 6000\text{ molecules})$. Nuclear decay energies are on the order of 1 MeV (1,000,000 eV) per event and can thus produce significant biological damage.

Conservation of Energy

The total energy of a system is conserved if there is no net addition (or subtraction) due to work or heat transfer. For conservative forces, such as the electrostatic force, conservation of energy states that mechanical energy is a constant.

Mechanical energy is the sum of the kinetic energy and potential energy of a system; that is, $K+U=\text{constant}$. A loss of U for a charged particle becomes an increase in its K. Conservation of energy is stated in equation form as

Equation:

$$K+U=\text{constant}$$

or

Equation:

$$K_{\text{i}}+U_{\text{i}}=K_{\text{f}}+U_{\text{f}}$$

where i and f stand for initial and final conditions. As we have found many times before, considering energy can give us insights and facilitate problem solving.

Example:

Electrical Potential Energy Converted into Kinetic Energy

Calculate the final speed of a free electron accelerated from rest through a potential difference of 100 V. (Assume that this numerical value is accurate to three significant figures.)

Strategy

We have a system with only conservative forces. Assuming the electron is accelerated in a vacuum, and neglecting the gravitational force (we will check on this assumption later), all of the electrical potential energy is converted into kinetic energy. We can identify the initial and final forms of energy to be $K_{\text{i}}=0, K_{\text{f}}=\frac{1}{2}mv^2, U_{\text{i}}=qV, U_{\text{f}}=0.$

Solution

Conservation of energy states that

Equation:

$$K_{\text{i}}+U_{\text{i}}=K_{\text{f}}+U_{\text{f}}.$$

Entering the forms identified above, we obtain

Equation:

$$qV=\frac{mv^2}{2}.$$

We solve this for *v*:

Equation:

$$v=\sqrt{\frac{2qV}{m}}.$$

Entering values for q, V, and m gives

Equation:

$$v=\sqrt{\frac{2(-1.60\ \times\ 10^{-19}\ \text{C})(-100\ \text{J/C})}{9.11\ \times\ 10^{-31}\ \text{kg}}}=5.93\ \times\ 10^{6}\ \text{m/s}.$$

Significance

Note that both the charge and the initial voltage are negative, as in [link]. From the discussion of electric charge and electric field, we know that electrostatic forces on small particles are generally very large compared with the gravitational force. The large final speed confirms that the gravitational force is indeed negligible here. The large speed also indicates how easy it is to accelerate electrons with small voltages because of their very small mass. Voltages much higher than the 100 V in this problem are typically used in electron guns. These higher voltages produce electron speeds so great that effects from special relativity must be taken into account and hence are reserved for a later chapter (Relativity). That is why we consider a low voltage (accurately) in this example.

Note:

Exercise:

Problem:

Check Your Understanding How would this example change with a positron? A positron is identical to an electron except the charge is positive.

Solution:

It would be going in the opposite direction, with no effect on the calculations as presented.

Voltage and Electric Field

So far, we have explored the relationship between voltage and energy. Now we want to explore the relationship between voltage and electric field. We will start with the general case for a non-uniform $\vec{\mathbf{E}}$ field. Recall that our general formula for the potential energy of a test charge q at point P relative to reference point R is

Equation:

$$U_P=-\int_R^P \vec{\mathbf{F}}\cdot d\vec{\mathbf{l}}.$$

When we substitute in the definition of electric field $(\vec{\mathbf{E}} = \vec{\mathbf{F}}/q)$, this becomes
Equation:

$$U_P = -q \int_R^P \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}.$$

Applying our definition of potential $(V = U/q)$ to this potential energy, we find that, in general,

Note:
Equation:

$$V_P = -\int_R^P \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}.$$

From our previous discussion of the potential energy of a charge in an electric field, the result is independent of the path chosen, and hence we can pick the integral path that is most convenient.

Consider the special case of a positive point charge q at the origin. To calculate the potential caused by q at a distance r from the origin relative to a reference of 0 at infinity (recall that we did the same for potential energy), let $P = r$ and $R = \infty$, with $d\vec{\mathbf{l}} = d\vec{\mathbf{r}} = \hat{\mathbf{r}}dr$ and use $\overrightarrow{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$. When we evaluate the integral
Equation:

$$V_P = -\int_R^P \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}$$

for this system, we have
Equation:

$$V_r = -\int_\infty^r \frac{kq}{r^2}\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}dr,$$

which simplifies to

Equation:

$$V_r = -\int_{\infty}^{r} \frac{kq}{r^2}\, dr = \frac{kq}{r} - \frac{kq}{\infty} = \frac{kq}{r}.$$

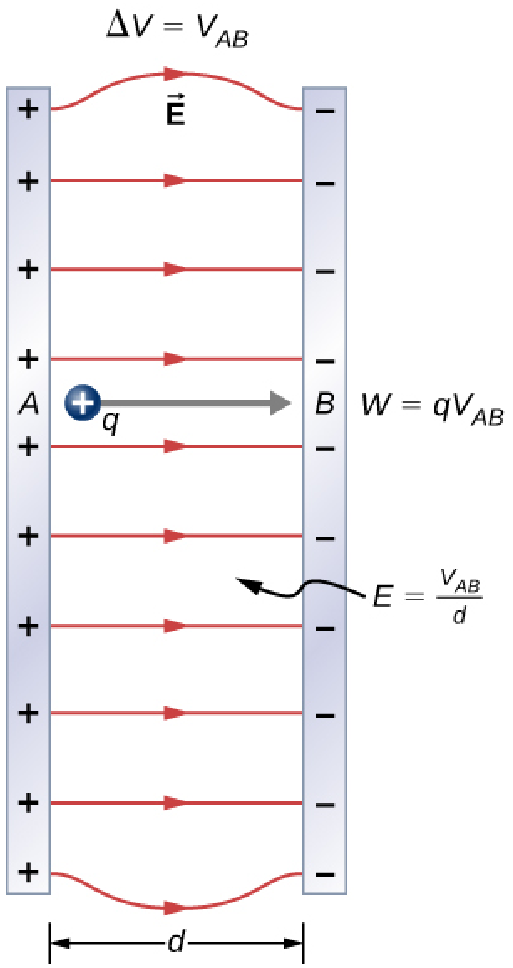
This result,

Equation:

$$V_r = \frac{kq}{r}$$

is the standard form of the potential of a point charge. This will be explored further in the next section.

To examine another interesting special case, suppose a uniform electric field $\vec{\mathbf{E}}$ is produced by placing a potential difference (or voltage) ΔV across two parallel metal plates, labeled *A* and *B* ([link]). Examining this situation will tell us what voltage is needed to produce a certain electric field strength. It will also reveal a more fundamental relationship between electric potential and electric field.



The relationship between V and E for parallel conducting plates is $E = V/d$. (Note that $\Delta V = V_{AB}$ in magnitude. For a charge that is moved from plate A at higher potential to plate B at lower potential, a minus sign needs to be included as follows: $-\Delta V = V_A - V_B = V_{AB}$.)

From a physicist's point of view, either ΔV or $\vec{\mathbf{E}}$ can be used to describe any interaction between charges. However, ΔV is a scalar quantity and has no direction, whereas $\vec{\mathbf{E}}$ is a vector quantity, having both magnitude and direction. (Note that the magnitude of the electric field, a scalar quantity, is represented by E.) The relationship between ΔV and $\vec{\mathbf{E}}$

is revealed by calculating the work done by the electric force in moving a charge from point A to point B. But, as noted earlier, arbitrary charge distributions require calculus. We therefore look at a uniform electric field as an interesting special case.

The work done by the electric field in [link] to move a positive charge q from A, the positive plate, higher potential, to B, the negative plate, lower potential, is

Equation:

$$W = -\Delta U = -q\Delta V.$$

The potential difference between points A and B is

Equation:

$$-\Delta V = -(V_B - V_A) = V_A - V_B = V_{AB}.$$

Entering this into the expression for work yields

Equation:

$$W = qV_{AB}.$$

Work is $W = \overrightarrow{\mathbf{F}} \cdot \vec{\mathbf{d}} = Fd\cos\theta$; here $\cos\theta = 1$, since the path is parallel to the field. Thus, $W = Fd$. Since $F = qE$, we see that $W = qEd$.

Substituting this expression for work into the previous equation gives

Equation:

$$qEd = qV_{AB}.$$

The charge cancels, so we obtain for the voltage between points A and B

Equation:

$$\left.\begin{array}{l} V_{AB} = Ed \\ E = \frac{V_{AB}}{d} \end{array}\right\} (\text{uniform } E\text{-field only})$$

where d is the distance from A to B, or the distance between the plates in [link]. Note that this equation implies that the units for electric field are volts per meter. We already know the units for electric field are newtons per coulomb; thus, the following relation among units is valid:

Equation:

$$1\,\text{N} / \text{C} = 1\,\text{V} / \text{m}.$$

Furthermore, we may extend this to the integral form. Substituting [link] into our definition for the potential difference between points *A* and *B*, we obtain

Equation:

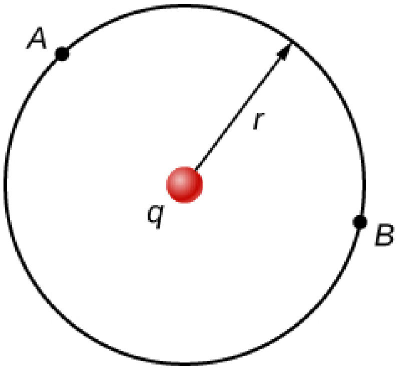
$$V_{BA} = V_B - V_A = -\int_R^B \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}} + \int_R^A \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}$$

which simplifies to

Equation:

$$V_B - V_A = -\int_A^B \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}.$$

As a demonstration, from this we may calculate the potential difference between two points (*A* and *B*) equidistant from a point charge *q* at the origin, as shown in [link].



The arc for calculating the potential difference between two points that are equidistant from a point charge at the origin.

To do this, we integrate around an arc of the circle of constant radius r between *A* and *B*, which means we let $d\vec{\mathbf{l}} = r\widehat{\varphi}d\varphi$, while using $\overrightarrow{\mathbf{E}} = \frac{kq}{r^2}\hat{\mathbf{r}}$. Thus,

Note:
Equation:

$$\Delta V_{BA} = V_B - V_A = -\int_A^B \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}$$

for this system becomes
Equation:

$$V_B - V_A = -\int_A^B \frac{kq}{r^2} \hat{\mathbf{r}} \cdot r\widehat{\boldsymbol{\varphi}} d\varphi.$$

However, $\hat{\mathbf{r}} \cdot \widehat{\boldsymbol{\varphi}} = 0$ and therefore
Equation:

$$V_B - V_A = 0.$$

This result, that there is no difference in potential along a constant radius from a point charge, will come in handy when we map potentials.

Example:
What Is the Highest Voltage Possible between Two Plates?
Dry air can support a maximum electric field strength of about $3.0 \times 10^6 \text{ V/m}$. Above that value, the field creates enough ionization in the air to make the air a conductor. This allows a discharge or spark that reduces the field. What, then, is the maximum voltage between two parallel conducting plates separated by 2.5 cm of dry air?
Strategy
We are given the maximum electric field *E* between the plates and the distance *d* between them. We can use the equation $V_{AB} = Ed$ to calculate the maximum voltage.
Solution
The potential difference or voltage between the plates is
Equation:

$$V_{AB} = Ed.$$

Entering the given values for *E* and *d* gives
Equation:

$$V_{AB} = (3.0 \times 10^6\,\text{V/m})(0.025\,\text{m}) = 7.5 \times 10^4\,\text{V}$$

or

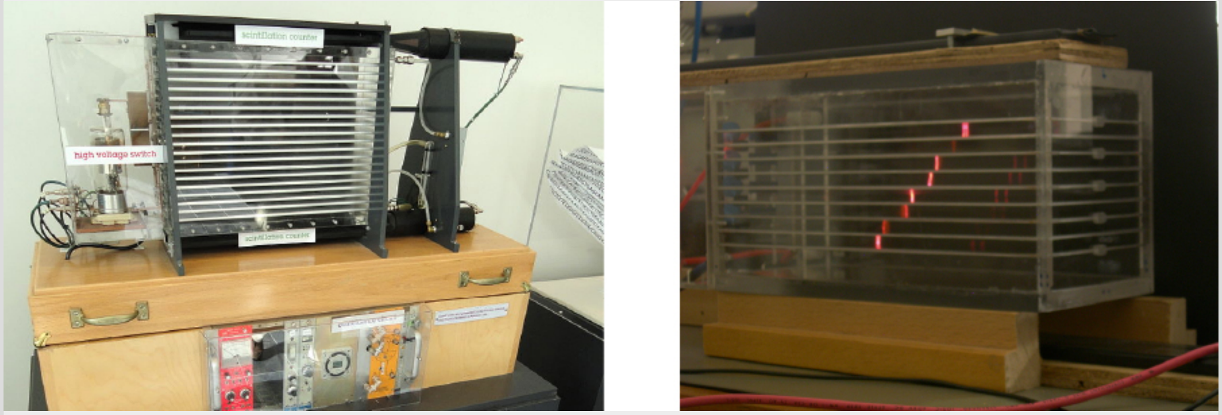
Equation:

$$V_{AB} = 75\,\text{kV}.$$

(The answer is quoted to only two digits, since the maximum field strength is approximate.)

Significance

One of the implications of this result is that it takes about 75 kV to make a spark jump across a 2.5-cm (1-in.) gap, or 150 kV for a 5-cm spark. This limits the voltages that can exist between conductors, perhaps on a power transmission line. A smaller voltage can cause a spark if there are spines on the surface, since sharp points have larger field strengths than smooth surfaces. Humid air breaks down at a lower field strength, meaning that a smaller voltage will make a spark jump through humid air. The largest voltages can be built up with static electricity on dry days ([link]).



A spark chamber is used to trace the paths of high-energy particles. Ionization created by the particles as they pass through the gas between the plates allows a spark to jump. The sparks are perpendicular to the plates, following electric field lines between them. The potential difference between adjacent plates is not high enough to cause sparks without the ionization produced by particles from accelerator experiments (or cosmic rays). This form of detector is now archaic and no longer in use except for demonstration purposes. (credit b: modification of work by Jack Collins)

Example:

Field and Force inside an Electron Gun

An electron gun ([link]) has parallel plates separated by 4.00 cm and gives electrons 25.0 keV of energy. (a) What is the electric field strength between the plates? (b) What force would this field exert on a piece of plastic with a $0.500\text{-}\mu\text{C}$ charge that gets between the plates?

Strategy

Since the voltage and plate separation are given, the electric field strength can be calculated directly from the expression $E=\frac{V_{AB}}{d}$. Once we know the electric field strength, we can find the force on a charge by using $\vec{\mathbf{F}}=q\vec{\mathbf{E}}$. Since the electric field is in only one direction, we can write this equation in terms of the magnitudes, $F=qE$.

Solution

a. The expression for the magnitude of the electric field between two uniform metal plates is
   Equation:

$$E=\frac{V_{AB}}{d}.$$

   Since the electron is a single charge and is given 25.0 keV of energy, the potential difference must be 25.0 kV. Entering this value for V_{AB} and the plate separation of 0.0400 m, we obtain
   Equation:

$$E=\frac{25.0\,\text{kV}}{0.0400\,\text{m}}=6.25\ \times\ 10^{5}\ \text{V/m}.$$

b. The magnitude of the force on a charge in an electric field is obtained from the equation
   Equation:

$$F=qE.$$

   Substituting known values gives
   Equation:

$$F=(0.500\ \times\ 10^{-6}\ \text{C})(6.25\ \times\ 10^{5}\ \text{V/m})=0.313\,\text{N}.$$

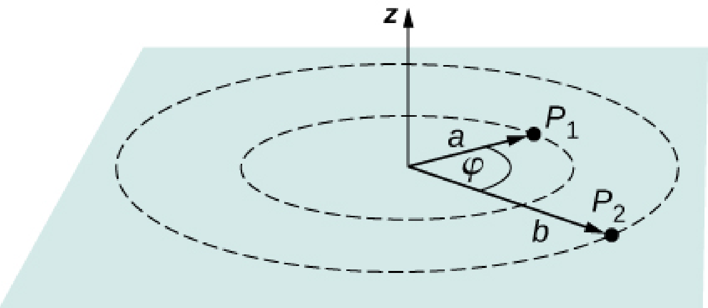
Significance

Note that the units are newtons, since $1\ \text{V/m}=1\,\text{N/C}$. Because the electric field is uniform between the plates, the force on the charge is the same no matter where the charge is located between the plates.

Example:

Calculating Potential of a Point Charge

Given a point charge $q=+2.0\,\text{nC}$ at the origin, calculate the potential difference between point P_1 a distance $a=4.0\,\text{cm}$ from q, and P_2 a distance $b=12.0\,\text{cm}$ from q, where the two points have an angle of $\varphi=24^{\circ}$ between them ([link]).



Find the difference in potential between P_1 and P_2.

Strategy

Do this in two steps. The first step is to use $V_B-V_A=-\int_A^B \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}}$ and let $A=a=4.0\,\text{cm}$ and $B=b=12.0\,\text{cm}$, with $d\vec{\mathbf{l}}=d\vec{\mathbf{r}}=\hat{\mathbf{r}}dr$ and $\vec{\mathbf{E}}=\frac{kq}{r^2}\hat{\mathbf{r}}$. Then perform the integral. The second step is to integrate $V_B-V_A=-\int_A^B \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}}$ around an arc of constant radius r, which means we let $d\vec{\mathbf{l}}=r\hat{\varphi}d\varphi$ with limits $0\le\varphi\le 24^{\circ}$, still using $\vec{\mathbf{E}}=\frac{kq}{r^2}\hat{\mathbf{r}}$. Then add the two results together.

Solution

For the first part, $V_B-V_A=-\int_A^B \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}}$ for this system becomes $V_b-V_a=-\int_a^b \frac{kq}{r^2}\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}dr$ which computes to

Equation:

$$\begin{aligned}\Delta V &= -\int_a^b \frac{kq}{r^2}dr = kq\left[\frac{1}{a}-\frac{1}{b}\right]\\ &= \left(8.99\times 10^9\,\text{Nm}^2/\text{C}^2\right)\left(2.0\times 10^{-9}\,\text{C}\right)\left[\frac{1}{0.040\,\text{m}}-\frac{1}{0.12\,\text{m}}\right]=300\,\text{V}.\end{aligned}$$

For the second step, $V_B - V_A = -\int_A^B \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}$ becomes $\Delta V = -\int_0^{24^\circ} \frac{kq}{r^2} \hat{\mathbf{r}} \cdot r\widehat{\boldsymbol{\varphi}} d\varphi$, but $\hat{\mathbf{r}} \cdot \widehat{\boldsymbol{\varphi}} = 0$ and therefore $\Delta V = 0$. Adding the two parts together, we get 300 V.

Significance

We have demonstrated the use of the integral form of the potential difference to obtain a numerical result. Notice that, in this particular system, we could have also used the formula for the potential due to a point charge at the two points and simply taken the difference.

Note:

Exercise:

Problem:

Check Your Understanding From the examples, how does the energy of a lightning strike vary with the height of the clouds from the ground? Consider the cloud-ground system to be two parallel plates.

Solution:

Given a fixed maximum electric field strength, the potential at which a strike occurs increases with increasing height above the ground. Hence, each electron will carry more energy. Determining if there is an effect on the total number of electrons lies in the future.

Before presenting problems involving electrostatics, we suggest a problem-solving strategy to follow for this topic.

Note:

Problem-Solving Strategy: Electrostatics

1. Examine the situation to determine if static electricity is involved; this may concern separated stationary charges, the forces among them, and the electric fields they create.
2. Identify the system of interest. This includes noting the number, locations, and types of charges involved.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful. Determine whether the Coulomb force is to be considered

directly—if so, it may be useful to draw a free-body diagram, using electric field lines.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns). It is important to distinguish the Coulomb force *F* from the electric field *E*, for example.
5. Solve the appropriate equation for the quantity to be determined (the unknown) or draw the field lines as requested.
6. Examine the answer to see if it is reasonable: Does it make sense? Are units correct and the numbers involved reasonable?

Summary

- Electric potential is potential energy per unit charge.
- The potential difference between points *A* and *B*, $V_B - V_A$, that is, the change in potential of a charge *q* moved from *A* to *B*, is equal to the change in potential energy divided by the charge.
- Potential difference is commonly called voltage, represented by the symbol ΔV: $\Delta V = \frac{\Delta U}{q}$ or $\Delta U = q\Delta V$.
- An electron-volt is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form, $1\text{ eV} = \left(1.60 \times 10^{-19}\text{ C}\right)(1\text{ V}) = \left(1.60 \times 10^{-19}\text{ C}\right)(1\text{ J/C}) = 1.60 \times 10^{-19}\text{ J}$.

Conceptual Questions

Exercise:

Problem:

Discuss how potential difference and electric field strength are related. Give an example.

Exercise:

Problem:

What is the strength of the electric field in a region where the electric potential is constant?

Solution:

The electric field strength is zero because electric potential differences are directly related to the field strength. If the potential difference is zero, then the field strength must also be zero.

Exercise:

Problem:

If a proton is released from rest in an electric field, will it move in the direction of increasing or decreasing potential? Also answer this question for an electron and a neutron. Explain why.

Exercise:

Problem:

Voltage is the common word for potential difference. Which term is more descriptive, voltage or potential difference?

Solution:

Potential difference is more descriptive because it indicates that it is the difference between the electric potential of two points.

Exercise:

Problem:

If the voltage between two points is zero, can a test charge be moved between them with zero net work being done? Can this necessarily be done without exerting a force? Explain.

Exercise:

Problem:

What is the relationship between voltage and energy? More precisely, what is the relationship between potential difference and electric potential energy?

Solution:

They are very similar, but potential difference is a feature of the system; when a charge is introduced to the system, it will have a potential energy which may be calculated by multiplying the magnitude of the charge by the potential difference.

Exercise:

Problem: Voltages are always measured between two points. Why?

Exercise:

Problem: How are units of volts and electron-volts related? How do they differ?

Solution:

An electron-volt is a volt multiplied by the charge of an electron. Volts measure potential difference, electron-volts are a unit of energy.

Exercise:

Problem:

Can a particle move in a direction of increasing electric potential, yet have its electric potential energy decrease? Explain

Problems

Exercise:

Problem:

Find the ratio of speeds of an electron and a negative hydrogen ion (one having an extra electron) accelerated through the same voltage, assuming non-relativistic final speeds. Take the mass of the hydrogen ion to be 1.67×10^{-27} kg.

Solution:

$$\begin{aligned} \frac{1}{2}m_e v_e^2 &= qV,\ \frac{1}{2}m_{\mathrm{H}}v_{\mathrm{H}}^2 = qV,\ \text{so that} \\ \frac{m_e v_e^2}{m_{\mathrm{H}}v_{\mathrm{H}}^2} &= 1 \text{ or } \frac{v_e}{v_{\mathrm{H}}} = 42.8 \end{aligned}$$

Exercise:

Problem:

An evacuated tube uses an accelerating voltage of 40 kV to accelerate electrons to hit a copper plate and produce X-rays. Non-relativistically, what would be the maximum speed of these electrons?

Exercise:

Problem:

Show that units of V/m and N/C for electric field strength are indeed equivalent.

Solution:

$$1\ \mathrm{V} = 1\ \mathrm{J/C};\ 1\ \mathrm{J} = 1\ \mathrm{N}\cdot\mathrm{m} \rightarrow 1\ \mathrm{V/m} = 1\ \mathrm{N/C}$$

Exercise:

Problem:

What is the strength of the electric field between two parallel conducting plates separated by 1.00 cm and having a potential difference (voltage) between them of 1.50×10^4 V?

Exercise:

Problem:

The electric field strength between two parallel conducting plates separated by 4.00 cm is 7.50×10^4 V/m. (a) What is the potential difference between the plates? (b) The plate with the lowest potential is taken to be zero volts. What is the potential 1.00 cm from that plate and 3.00 cm from the other?

Solution:

a. $V_{AB} = 3.00\,\text{kV}$; b. $V_{AB} = 750\text{ V}$

Exercise:

Problem:

The voltage across a membrane forming a cell wall is 80.0 mV and the membrane is 9.00 nm thick. What is the electric field strength? (The value is surprisingly large, but correct.) You may assume a uniform electric field.

Exercise:

Problem:

Two parallel conducting plates are separated by 10.0 cm, and one of them is taken to be at zero volts. (a) What is the electric field strength between them, if the potential 8.00 cm from the zero volt plate (and 2.00 cm from the other) is 450 V? (b) What is the voltage between the plates?

Solution:

a. $V_{AB} = Ed \rightarrow E = 5.63\,\text{kV/m}$;
b. $V_{AB} = 563\text{ V}$

Exercise:

Problem:

Find the maximum potential difference between two parallel conducting plates separated by 0.500 cm of air, given the maximum sustainable electric field strength in air to be 3.0×10^6 V/m.

Exercise:

Problem:

An electron is to be accelerated in a uniform electric field having a strength of $2.00\ \times\ 10^6\ \mathrm{V/m}$. (a) What energy in keV is given to the electron if it is accelerated through 0.400 m? (b) Over what distance would it have to be accelerated to increase its energy by 50.0 GeV?

Solution:

a. $$\begin{aligned}\Delta K &= q\Delta V \text{ and } V_{AB}=Ed, \text{ so that}\\ \Delta K &= 800\,\mathrm{keV};\end{aligned}$$

b. $d=25.0\,\mathrm{km}$

Exercise:

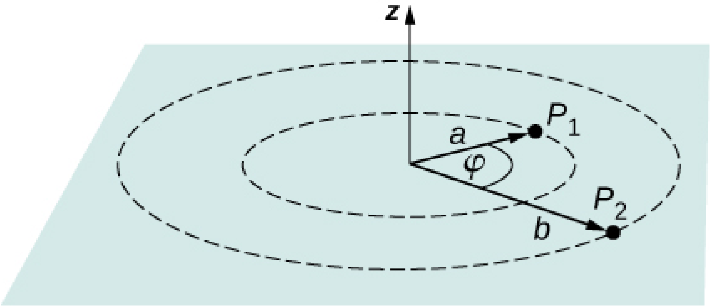
Problem:

Use the definition of potential difference in terms of electric field to deduce the formula for potential difference between $r=r_a$ and $r=r_b$ for a point charge located at the origin. Here *r* is the spherical radial coordinate.

Exercise:

Problem:

The electric field in a region is pointed away from the z-axis and the magnitude depends upon the distance *s* from the axis. The magnitude of the electric field is given as $E=\frac{\alpha}{s}$ where α is a constant. Find the potential difference between points P_1 and P_2, explicitly stating the path over which you conduct the integration for the line integral.



Solution:

One possibility is to stay at constant radius and go along the arc from P_1 to P_2, which will have zero potential due to the path being perpendicular to the electric field. Then integrate from *a* to *b*: $V_{ab}=\alpha\ln\left(\frac{b}{a}\right)$

Exercise:

Problem:

Singly charged gas ions are accelerated from rest through a voltage of 13.0 V. At what temperature will the average kinetic energy of gas molecules be the same as that given these ions?

Glossary

electric potential
potential energy per unit charge

electric potential difference
the change in potential energy of a charge *q* moved between two points, divided by the charge.

electron-volt
energy given to a fundamental charge accelerated through a potential difference of one volt

voltage
change in potential energy of a charge moved from one point to another, divided by the charge; units of potential difference are joules per coulomb, known as volt

Calculations of Electric Potential

By the end of this section, you will be able to:

- Calculate the potential due to a point charge
- Calculate the potential of a system of multiple point charges
- Describe an electric dipole
- Define dipole moment
- Calculate the potential of a continuous charge distribution

Point charges, such as electrons, are among the fundamental building blocks of matter. Furthermore, spherical charge distributions (such as charge on a metal sphere) create external electric fields exactly like a point charge. The electric potential due to a point charge is, thus, a case we need to consider.

We can use calculus to find the work needed to move a test charge *q* from a large distance away to a distance of *r* from a point charge *q*. Noting the connection between work and potential $W = -q\Delta V$, as in the last section, we can obtain the following result.

Note:

Electric Potential *V* of a Point Charge

The electric potential *V* of a point charge is given by

Equation:

$$V = \frac{kq}{r} \text{(point charge)}$$

where *k* is a constant equal to $8.99 \ \times \ 10^9 \ \mathrm{N \cdot m^2/C^2}$.

The potential at infinity is chosen to be zero. Thus, *V* for a point charge decreases with distance, whereas $\vec{\mathbf{E}}$ for a point charge decreases with distance squared:

Equation:

$$E = \frac{F}{q_t} = \frac{kq}{r^2}.$$

Recall that the electric potential *V* is a scalar and has no direction, whereas the electric field $\vec{\mathbf{E}}$ is a vector. To find the voltage due to a combination of point charges, you add the individual voltages as numbers. To find the total electric field, you must add the individual fields as vectors, taking magnitude and direction into account. This is consistent with the fact that *V* is closely associated with energy, a scalar, whereas $\vec{\mathbf{E}}$ is closely associated with force, a vector.

Example:

What Voltage Is Produced by a Small Charge on a Metal Sphere?

Charges in static electricity are typically in the nanocoulomb (nC) to microcoulomb (μC) range. What is the voltage 5.00 cm away from the center of a 1-cm-diameter solid metal sphere that has a –3.00-nC static charge?

Strategy

As we discussed in Electric Charges and Fields, charge on a metal sphere spreads out uniformly and produces a field like that of a point charge located at its center. Thus, we can find the voltage using the equation $V = \frac{kq}{r}$.

Solution

Entering known values into the expression for the potential of a point charge, we obtain

Equation:

$$V = k\frac{q}{r} = \left(8.99 \times 10^{9}\,\text{N}\cdot\text{m}^{2}/\text{C}^{2}\right)\left(\frac{-3.00 \times 10^{-9}\,\text{C}}{5.00 \times 10^{-2}\,\text{m}}\right) = -539\,\text{V}.$$

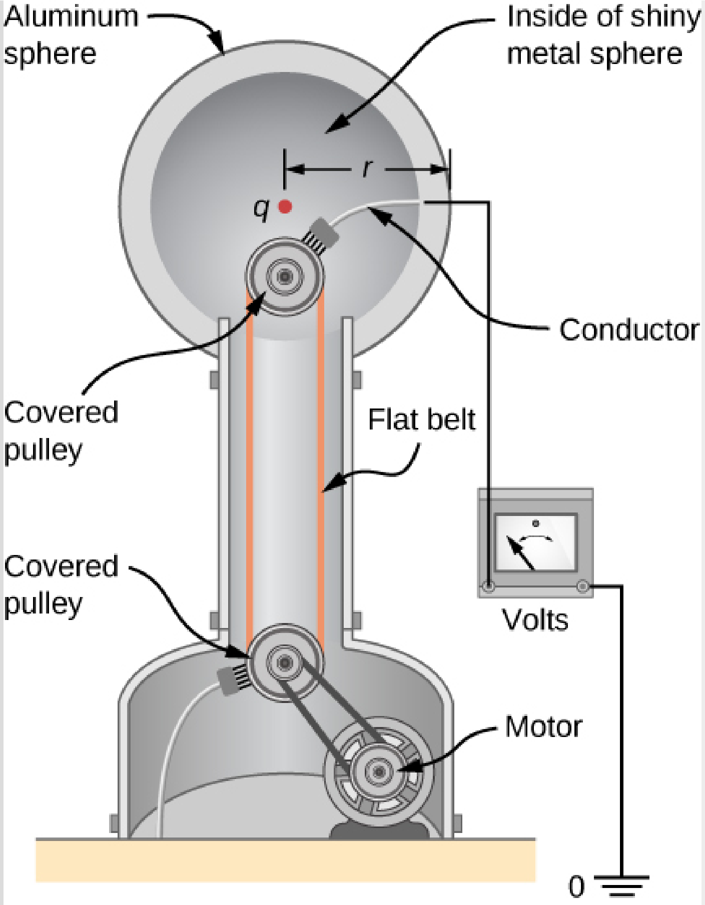
Significance

The negative value for voltage means a positive charge would be attracted from a larger distance, since the potential is lower (more negative) than at larger distances. Conversely, a negative charge would be repelled, as expected.

Example:

What Is the Excess Charge on a Van de Graaff Generator?

A demonstration Van de Graaff generator has a 25.0-cm-diameter metal sphere that produces a voltage of 100 kV near its surface ([link]). What excess charge resides on the sphere? (Assume that each numerical value here is shown with three significant figures.)


Aluminum sphere
Inside of shiny metal sphere
r
q
Conductor
Covered pulley
Flat belt
Covered pulley
Volts
Motor
0

The voltage of this demonstration Van de Graaff generator is measured between the charged sphere and ground. Earth’s potential is taken to be zero as a reference. The potential of the charged conducting sphere is the same as that of an equal point charge at its center.

Strategy

The potential on the surface is the same as that of a point charge at the center of the sphere, 12.5 cm away. (The radius of the sphere is 12.5 cm.) We can thus determine the excess charge using the equation

Equation:

$$V = \frac{kq}{r}.$$

Solution

Solving for q and entering known values gives

Equation:

$$q=\frac{rV}{k}=\frac{(0.125\,\text{m})\left(100\;\times\;10^3\,\text{V}\right)}{8.99\;\times\;10^9\,\text{N}\cdot\text{m}^2/\text{C}^2}=1.39\;\times\;10^{-6}\,\text{C}=1.39\,\mu\text{C}.$$

Significance
This is a relatively small charge, but it produces a rather large voltage. We have another indication here that it is difficult to store isolated charges.

Note:
Exercise:

Problem:

Check Your Understanding What is the potential inside the metal sphere in [link]?

Solution:

$V=k\frac{q}{r}=\left(8.99\;\times\;10^9\,\text{N}\cdot\text{m}^2/\text{C}^2\right)\left(\frac{-3.00\times10^{-9}\,\text{C}}{5.00\times10^{-3}\,\text{m}}\right)=-5390\,\text{V}$; recall that the electric field inside a conductor is zero. Hence, any path from a point on the surface to any point in the interior will have an integrand of zero when calculating the change in potential, and thus the potential in the interior of the sphere is identical to that on the surface.

The voltages in both of these examples could be measured with a meter that compares the measured potential with ground potential. Ground potential is often taken to be zero (instead of taking the potential at infinity to be zero). It is the potential difference between two points that is of importance, and very often there is a tacit assumption that some reference point, such as Earth or a very distant point, is at zero potential. As noted earlier, this is analogous to taking sea level as $h=0$ when considering gravitational potential energy $U_g=mgh$.

Systems of Multiple Point Charges

Just as the electric field obeys a superposition principle, so does the electric potential. Consider a system consisting of *N* charges $q_1,q_2,\ldots,q_N$. What is the net electric potential *V* at a space point *P* from these charges? Each of these charges is a source charge that produces its own electric potential at point *P*, independent of whatever other changes may be doing. Let $V_1,V_2,\ldots,V_N$ be the electric potentials at *P* produced by the charges $q_1,q_2,\ldots,q_N$, respectively. Then, the net electric potential V_P at that point is equal to the sum of these individual electric potentials. You can easily show this by calculating the potential energy of a test charge when you bring the test charge from the reference point at infinity to point *P*:
Equation:

$$V_P = V_1 + V_2 + \cdots + V_N = \sum_1^N V_i.$$

Note that electric potential follows the same principle of superposition as electric field and electric potential energy. To show this more explicitly, note that a test charge q_i at the point P in space has distances of $r_1, r_2, \ldots, r_N$ from the N charges fixed in space above, as shown in [link]. Using our formula for the potential of a point charge for each of these (assumed to be point) charges, we find that

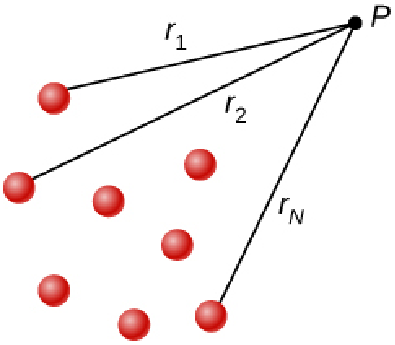
Note:
Equation:

$$V_P = \sum_1^N k\frac{q_i}{r_i} = k\sum_1^N \frac{q_i}{r_i}.$$

Therefore, the electric potential energy of the test charge is
Equation:

$$U_P = q_t V_P = q_t k \sum_1^N \frac{q_i}{r_i},$$

which is the same as the work to bring the test charge into the system, as found in the first section of the chapter.



Notation for direct distances from charges to a space point *P*.

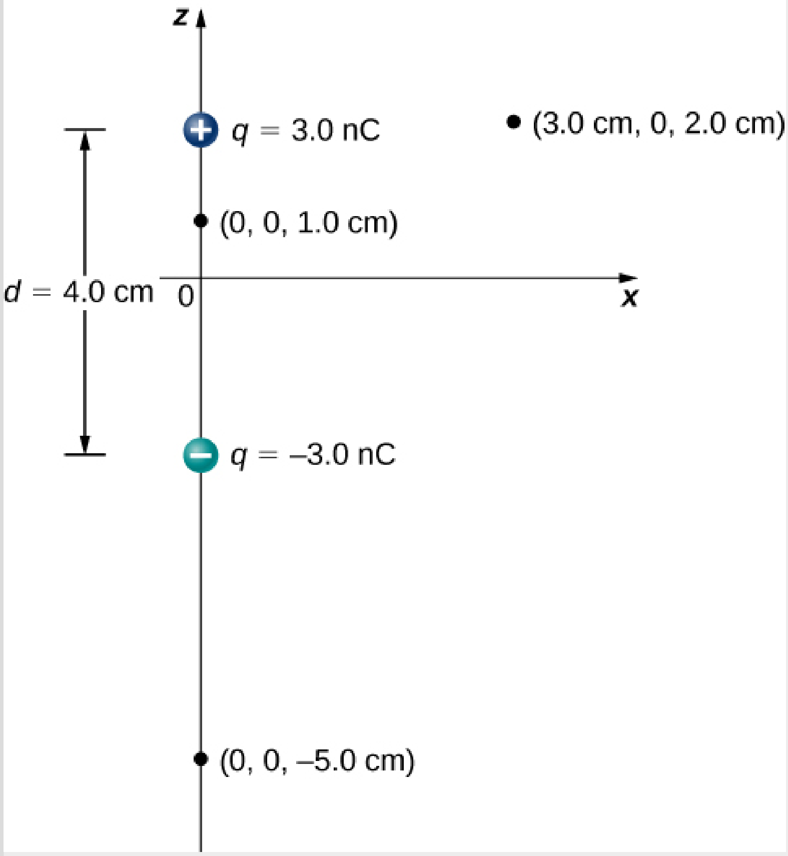
The Electric Dipole

An **electric dipole** is a system of two equal but opposite charges a fixed distance apart. This system is used to model many real-world systems, including atomic and molecular interactions. One of these systems is the water molecule, under certain circumstances. These circumstances are met inside a microwave oven, where electric fields with alternating directions make the water molecules change orientation. This vibration is the same as heat at the molecular level.

Example:

Electric Potential of a Dipole

Consider the dipole in [link] with the charge magnitude of $q = 3.0\,\text{nC}$ and separation distance $d = 4.0\,\text{cm}$. What is the potential at the following locations in space? (a) (0, 0, 1.0 cm); (b) (0, 0, –5.0 cm); (c) (3.0 cm, 0, 2.0 cm).



A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point *P* in space.

Strategy

Apply $V_P = k\sum_1^N \frac{q_i}{r_i}$ to each of these three points.

Solution

a. $V_P = k\sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9\,\mathrm{N \cdot m^2/C^2})\left(\frac{3.0\,\mathrm{nC}}{0.010\,\mathrm{m}} - \frac{3.0\,\mathrm{nC}}{0.030\,\mathrm{m}}\right) = 1.8 \times 10^3\,\mathrm{V}$

b. $V_P = k\sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9\,\mathrm{N \cdot m^2/C^2})\left(\frac{3.0\,\mathrm{nC}}{0.070\,\mathrm{m}} - \frac{3.0\,\mathrm{nC}}{0.030\,\mathrm{m}}\right) = -5.1 \times 10^2\,\mathrm{V}$

c. $V_P = k\sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9\,\mathrm{N \cdot m^2/C^2})\left(\frac{3.0\,\mathrm{nC}}{0.030\,\mathrm{m}} - \frac{3.0\,\mathrm{nC}}{0.050\,\mathrm{m}}\right) = 3.6 \times 10^2\,\mathrm{V}$

Significance

Note that evaluating potential is significantly simpler than electric field, due to potential being a scalar instead of a vector.

Note:

Exercise:

Problem: Check Your Understanding What is the potential on the *x*-axis? The *z*-axis?

Solution:

The *x*-axis the potential is zero, due to the equal and opposite charges the same distance from it. On the *z*-axis, we may superimpose the two potentials; we will find that for $z >> d$, again the potential goes to zero due to cancellation.

Now let us consider the special case when the distance of the point *P* from the dipole is much greater than the distance between the charges in the dipole, $r \gg d$; for example, when we are interested in the electric potential due to a polarized molecule such as a water molecule. This is not so far (infinity) that we can simply treat the potential as zero, but the distance is great enough that we can simplify our calculations relative to the previous example.

We start by noting that in [link] the potential is given by

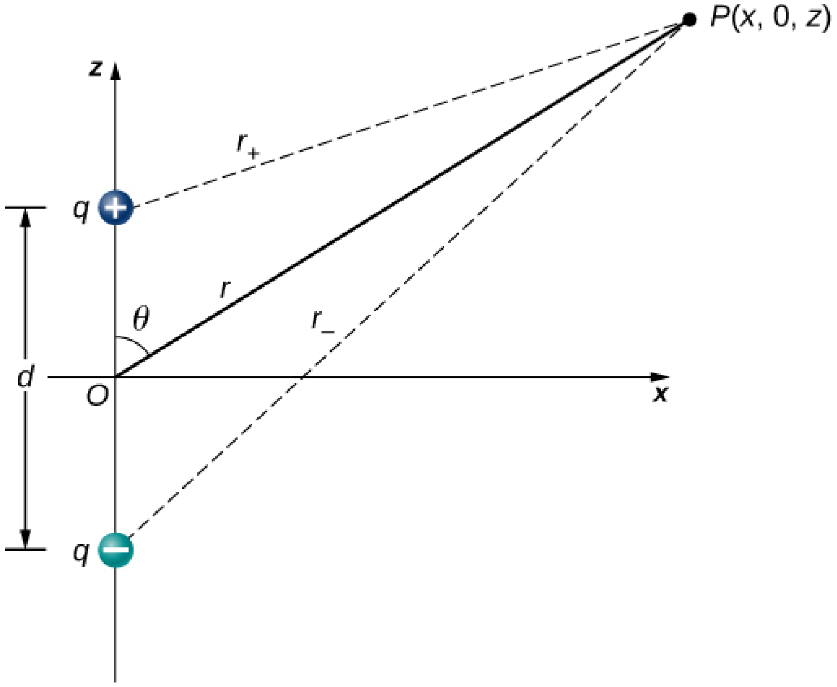
Equation:

$$V_P = V_+ + V_- = k\left(\frac{q}{r_+} - \frac{q}{r_-}\right)$$

where

Equation:

$$r_\pm = \sqrt{x^2 + \left(z \mp \frac{d}{2}\right)^2}.$$



A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point *P* in space.

This is still the exact formula. To take advantage of the fact that $r \gg d$, we rewrite the radii in terms of polar coordinates, with $x = r \sin\theta$ and $z = r\cos\theta$. This gives us

Equation:

$$r_\pm = \sqrt{r^2 \sin^2\theta + \left(r\cos\theta \mp \frac{d}{2}\right)^2}.$$

We can simplify this expression by pulling *r* out of the root,
Equation:

$$r_{\pm} = r\sqrt{\sin^2\theta + \left(\cos\theta \mp \frac{d}{2r}\right)^2}$$

and then multiplying out the parentheses
Equation:

$$r_{\pm} = r\sqrt{\sin^2\theta + \cos^2\theta \mp \cos\theta\frac{d}{r} + \left(\frac{d}{2r}\right)^2} = r\sqrt{1 \mp \cos\theta\frac{d}{r} + \left(\frac{d}{2r}\right)^2}.$$

The last term in the root is small enough to be negligible (remember $r \gg d$, and hence $(d/r)^2$ is extremely small, effectively zero to the level we will probably be measuring), leaving us with
Equation:

$$r_{\pm} = r\sqrt{1 \mp \cos\theta\frac{d}{r}}.$$

Using the binomial approximation (a standard result from the mathematics of series, when α is small)
Equation:

$$\frac{1}{\sqrt{1 \mp \alpha}} \approx 1 \pm \frac{\alpha}{2}$$

and substituting this into our formula for V_P , we get
Equation:

$$V_P = k\left[\frac{q}{r}\left(1 + \frac{d\cos\theta}{2r}\right) - \frac{q}{r}\left(1 - \frac{d\cos\theta}{2r}\right)\right] = k\frac{qd\cos\theta}{r^2}.$$

This may be written more conveniently if we define a new quantity, the **electric dipole moment**,

Note:
Equation:

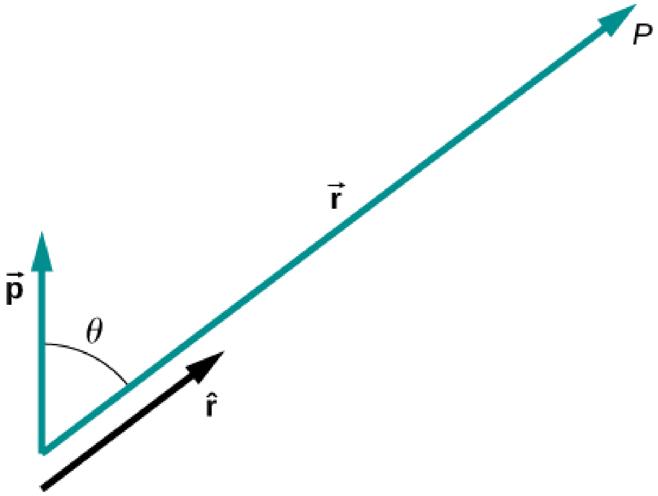
$$\vec{\mathbf{p}} = q\vec{\mathbf{d}},$$

where these vectors point from the negative to the positive charge. Note that this has magnitude *qd*. This quantity allows us to write the potential at point *P* due to a dipole at the origin as

Note:
Equation:

$$V_P = k\frac{\vec{\mathbf{p}} \cdot \hat{\mathbf{r}}}{r^2}.$$

A diagram of the application of this formula is shown in [link].



The geometry for the application of the potential of a dipole.

There are also higher-order moments, for quadrupoles, octupoles, and so on. You will see these in future classes.

Potential of Continuous Charge Distributions

We have been working with point charges a great deal, but what about continuous charge distributions? Recall from [link] that

Equation:

$$V_P = k\sum \frac{q_i}{r_i}.$$

We may treat a continuous charge distribution as a collection of infinitesimally separated individual points. This yields the integral

Note:

Equation:

$$V_P = k\int \frac{dq}{r}$$

for the potential at a point P. Note that r is the distance from each individual point in the charge distribution to the point P. As we saw in Electric Charges and Fields, the infinitesimal charges are given by

Equation:

$$dq = \begin{cases} \lambda\, dl & \text{(one dimension)} \\ \sigma\, dA & \text{(two dimensions)} \\ \rho\, dV & \text{(three dimensions)} \end{cases}$$

where λ is linear charge density, σ is the charge per unit area, and ρ is the charge per unit volume.

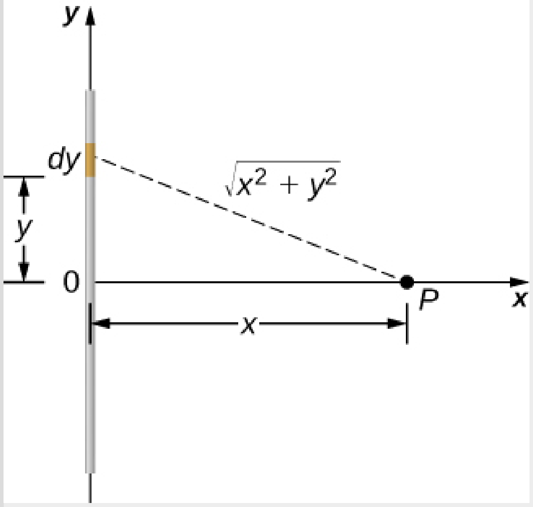
Example:

Potential of a Line of Charge

Find the electric potential of a uniformly charged, nonconducting wire with linear density λ (coulomb/meter) and length L at a point that lies on a line that divides the wire into two equal parts.

Strategy

To set up the problem, we choose Cartesian coordinates in such a way as to exploit the symmetry in the problem as much as possible. We place the origin at the center of the wire and orient the *y*-axis along the wire so that the ends of the wire are at $y = \pm L/2$. The field point *P* is in the *xy*-plane and since the choice of axes is up to us, we choose the *x*-axis to pass through the field point *P*, as shown in [link].



We want to calculate the electric potential due to a line of charge.

Solution

Consider a small element of the charge distribution between *y* and $y + dy$. The charge in this cell is $dq = \lambda\, dy$ and the distance from the cell to the field point *P* is $\sqrt{x^2 + y^2}$. Therefore, the potential becomes

Equation:

$$\begin{aligned} V_P &= k\int \frac{dq}{r} = k\int_{-L/2}^{L/2} \frac{\lambda dy}{\sqrt{x^2+y^2}} = k\lambda\left[\ln\left(y + \sqrt{y^2+x^2}\right)\right]_{-L/2}^{L/2} \\ &= k\lambda\left[\ln\left(\left(\tfrac{L}{2}\right) + \sqrt{\left(\tfrac{L}{2}\right)^2 + x^2}\right) - \ln\left(\left(-\tfrac{L}{2}\right) + \sqrt{\left(-\tfrac{L}{2}\right)^2 + x^2}\right)\right] \\ &= k\lambda \ln\left[\tfrac{L+\sqrt{L^2+4x^2}}{-L+\sqrt{L^2+4x^2}}\right]. \end{aligned}$$

Significance

Note that this was simpler than the equivalent problem for electric field, due to the use of scalar quantities. Recall that we expect the zero level of the potential to be at infinity, when we have a finite charge. To examine this, we take the limit of the above potential as *x* approaches infinity; in this case, the terms inside the natural log approach one, and hence the potential approaches zero in this limit. Note that we could have done this problem equivalently in cylindrical coordinates; the only effect would be to substitute *r* for *x* and *z* for *y*.

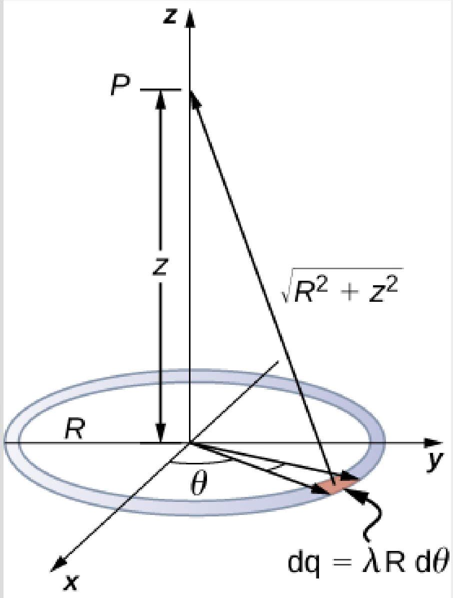
Example:

Potential Due to a Ring of Charge

A ring has a uniform charge density λ, with units of coulomb per unit meter of arc. Find the electric potential at a point on the axis passing through the center of the ring.

Strategy

We use the same procedure as for the charged wire. The difference here is that the charge is distributed on a circle. We divide the circle into infinitesimal elements shaped as arcs on the circle and use cylindrical coordinates shown in [link].



We want to calculate the electric potential due to a ring of charge.

Solution

A general element of the arc between θ and $\theta + d\theta$ is of length $Rd\theta$ and therefore contains a charge equal to $\lambda Rd\theta$. The element is at a distance of $\sqrt{z^2 + R^2}$ from P, and therefore the potential is

Equation:

$$V_P = k\int \frac{dq}{r} = k\int_0^{2\pi} \frac{\lambda Rd\theta}{\sqrt{z^2 + R^2}} = \frac{k\lambda R}{\sqrt{z^2 + R^2}} \int_0^{2\pi} d\theta = \frac{2\pi k\lambda R}{\sqrt{z^2 + R^2}} = k\frac{q_{\text{tot}}}{\sqrt{z^2 + R^2}}.$$

Significance

This result is expected because every element of the ring is at the same distance from point P. The net potential at P is that of the total charge placed at the common distance, $\sqrt{z^2 + R^2}$.

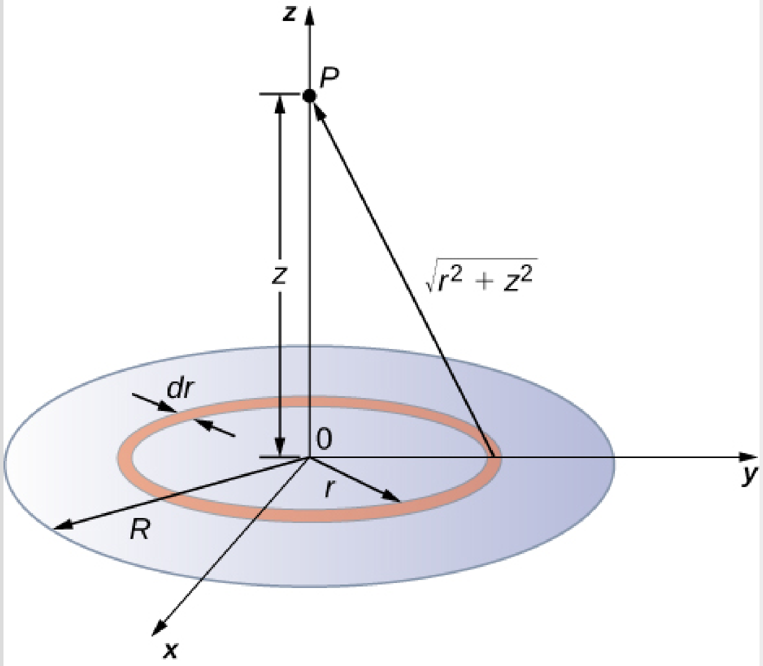
Example:

Potential Due to a Uniform Disk of Charge

A disk of radius R has a uniform charge density σ, with units of coulomb meter squared. Find the electric potential at any point on the axis passing through the center of the disk.

Strategy

We divide the disk into ring-shaped cells, and make use of the result for a ring worked out in the previous example, then integrate over r in addition to θ. This is shown in [link].



We want to calculate the electric potential due to a disk of charge.

Solution

An infinitesimal width cell between cylindrical coordinates r and $r + dr$ shown in [link] will be a ring of charges whose electric potential dV_P at the field point has the following expression

Equation:

$$dV_P = k\frac{dq}{\sqrt{z^2+r^2}}$$

where

Equation:

$$dq = \sigma 2\pi r dr.$$

The superposition of potential of all the infinitesimal rings that make up the disk gives the net potential at point P. This is accomplished by integrating from $r = 0$ to $r = R$:

Equation:

$$\begin{aligned} V_P &= \int dV_P = k2\pi\sigma \int_0^R \frac{r\,dr}{\sqrt{z^2+r^2}}, \\ &= k2\pi\sigma\left(\sqrt{z^2+R^2}-\sqrt{z^2}\right). \end{aligned}$$

Significance

The basic procedure for a disk is to first integrate around *θ* and then over *r*. This has been demonstrated for uniform (constant) charge density. Often, the charge density will vary with *r*, and then the last integral will give different results.

Example:

Potential Due to an Infinite Charged Wire

Find the electric potential due to an infinitely long uniformly charged wire.

Strategy

Since we have already worked out the potential of a finite wire of length *L* in [link], we might wonder if taking $L \to \infty$ in our previous result will work:

Equation:

$$V_P = \lim_{L\to\infty} k\lambda \ln\left(\frac{L+\sqrt{L^2+4x^2}}{-L+\sqrt{L^2+4x^2}}\right).$$

However, this limit does not exist because the argument of the logarithm becomes [2/0] as $L \to \infty$, so this way of finding *V* of an infinite wire does not work. The reason for this problem may be traced to the fact that the charges are not localized in some space but continue to infinity in the direction of the wire. Hence, our (unspoken) assumption that zero potential must be an infinite distance from the wire is no longer valid.

To avoid this difficulty in calculating limits, let us use the definition of potential by integrating over the electric field from the previous section, and the value of the electric field from this charge configuration from the previous chapter.

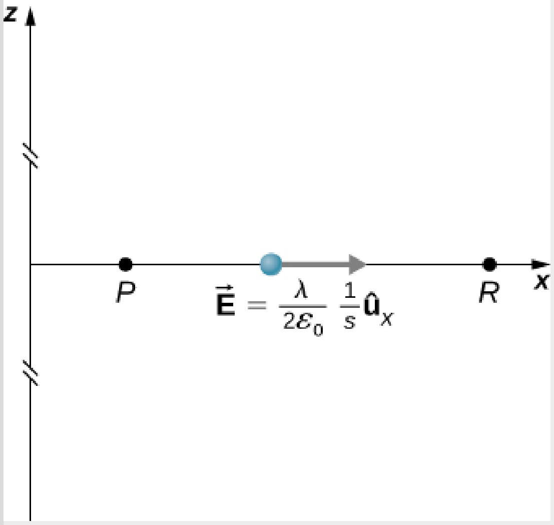
Solution

We use the integral

Equation:

$$V_P = -\int_R^P \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}}$$

where *R* is a finite distance from the line of charge, as shown in [link].


z
P
$\vec{\mathbf{E}} = \frac{\lambda}{2\varepsilon_0}\frac{1}{s}\hat{\mathbf{u}}_x$
R
x

Points of interest for calculating the potential of an infinite line of charge.

With this setup, we use $\overrightarrow{\mathbf{E}}_P = 2k\lambda\frac{1}{s}\hat{\mathbf{s}}$ and $d\vec{\mathbf{l}} = d\vec{\mathbf{s}}$ to obtain

Equation:

$$V_P - V_R = -\int_R^P 2k\lambda\frac{1}{s}ds = -2k\lambda \ln\frac{s_P}{s_R}.$$

Now, if we define the reference potential $V_R = 0$ at $s_R = 1\ \text{m}$, this simplifies to

Equation:

$$V_P = -2k\lambda\ln s_P.$$

Note that this form of the potential is quite usable; it is 0 at 1 m and is undefined at infinity, which is why we could not use the latter as a reference.

Significance

Although calculating potential directly can be quite convenient, we just found a system for which this strategy does not work well. In such cases, going back to the definition of potential in terms of the electric field may offer a way forward.

Note:

Exercise:

Problem:

Check Your Understanding What is the potential on the axis of a nonuniform ring of charge, where the charge density is $\lambda(\theta) = \lambda \cos\theta$?

Solution:

It will be zero, as at all points on the axis, there are equal and opposite charges equidistant from the point of interest. Note that this distribution will, in fact, have a dipole moment.

Summary

- Electric potential is a scalar whereas electric field is a vector.
- Addition of voltages as numbers gives the voltage due to a combination of point charges, allowing us to use the principle of superposition: $V_P = k \sum_1^N \frac{q_i}{r_i}$.
- An electric dipole consists of two equal and opposite charges a fixed distance apart, with a dipole moment $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$.
- Continuous charge distributions may be calculated with $V_P = k \int \frac{dq}{r}$.

Conceptual Questions

Exercise:

Problem:

Compare the electric dipole moments of charges $\pm Q$ separated by a distance *d* and charges $\pm Q/2$ separated by a distance *d*/2.

Solution:

The second has 1/4 the dipole moment of the first.

Exercise:

Problem:

Would Gauss’s law be helpful for determining the electric field of a dipole? Why?

Exercise:

Problem:

In what region of space is the potential due to a uniformly charged sphere the same as that of a point charge? In what region does it differ from that of a point charge?

Solution:

The region outside of the sphere will have a potential indistinguishable from a point charge; the interior of the sphere will have a different potential.

Exercise:

Problem:

Can the potential of a nonuniformly charged sphere be the same as that of a point charge? Explain.

Problems

Exercise:

Problem:

A 0.500-cm-diameter plastic sphere, used in a static electricity demonstration, has a uniformly distributed 40.0-pC charge on its surface. What is the potential near its surface?

Solution:

$V = 144\text{ V}$

Exercise:

Problem:

How far from a 1.00-μC point charge is the potential 100 V? At what distance is it 2.00×10^2 V?

Exercise:

Problem:

If the potential due to a point charge is 5.00×10^2 V at a distance of 15.0 m, what are the sign and magnitude of the charge?

Solution:

$V = \frac{kQ}{r} \rightarrow Q = 8.33 \times 10^{-7}\text{ C};$
The charge is positive because the potential is positive.

Exercise:

Problem:

In nuclear fission, a nucleus splits roughly in half. (a) What is the potential $2.00 \times 10^{-14}\,\text{m}$ from a fragment that has 46 protons in it? (b) What is the potential energy in MeV of a similarly charged fragment at this distance?

Exercise:

Problem:

A research Van de Graaff generator has a 2.00-m-diameter metal sphere with a charge of 5.00 mC on it. Assume the potential energy is zero at a reference point infinitely far away from the Van de Graaff. (a) What is the potential near its surface? (b) At what distance from its center is the potential 1.00 MV? (c) An oxygen atom with three missing electrons is released near the Van de Graaff generator. What is its kinetic energy in MeV when the atom is at the distance found in part b?

Solution:

a. $V = 45.0\,\text{MV}$;
b. $V = \frac{kQ}{r} \rightarrow r = 45.0\,\text{m}$;
c. $\Delta U = 132\,\text{MeV}$

Exercise:

Problem:

An electrostatic paint sprayer has a 0.200-m-diameter metal sphere at a potential of 25.0 kV that repels paint droplets onto a grounded object.

(a) What charge is on the sphere? (b) What charge must a 0.100-mg drop of paint have to arrive at the object with a speed of 10.0 m/s?

Exercise:

Problem:

(a) What is the potential between two points situated 10 cm and 20 cm from a $3.0\text{-}\mu\text{C}$ point charge? (b) To what location should the point at 20 cm be moved to increase this potential difference by a factor of two?

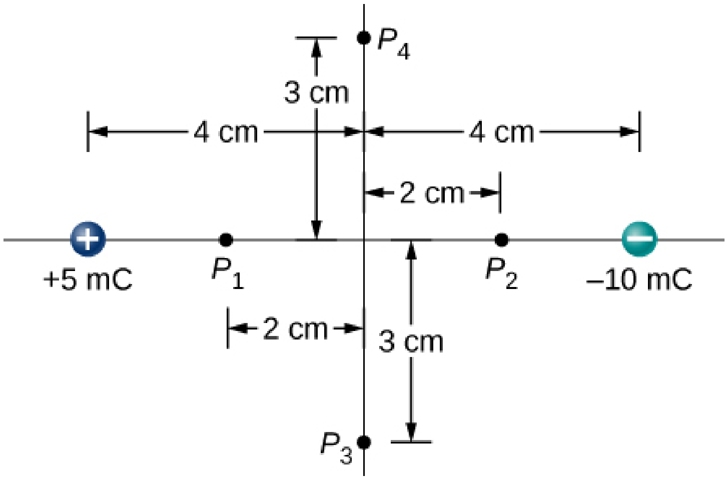
Solution:

$V = kQ/r$; a. Relative to origin, find the potential at each point and then calculate the difference.
$\Delta V = 135 \times 10^3$ V;
b. To double the potential difference, move the point from 20 cm to infinity; the potential at 20 cm is halfway between zero and that at 10 cm.

Exercise:

Problem:

Find the potential at points $P_1, P_2, P_3,$ and P_4 in the diagram due to the two given charges.



Exercise:

Problem:

Two charges –2.0 μC and $+2.0\ \mu$C are separated by 4.0 cm on the *z*-axis symmetrically about origin, with the positive one uppermost. Two space points of interest P_1 and P_2 are located 3.0 cm and 30 cm from origin at an angle 30° with respect to the *z*-axis. Evaluate electric potentials at P_1 and P_2 in two ways: (a) Using the exact formula for point charges, and (b) using the approximate dipole potential formula.

Solution:

a. $V_{P1} = 7.4 \times 10^5$ V
and $V_{P2} = 6.9 \times 10^3$ V;
b. $V_{P1} = 6.9 \times 10^5$ V and $V_{P2} = 6.9 \times 10^3$ V

Exercise:

Problem:

(a) Plot the potential of a uniformly charged 1-m rod with 1 C/m charge as a function of the perpendicular distance from the center. Draw your graph from $\mathrm{s} = 0.1\ \mathrm{m}$ to $\mathrm{s} = 1.0\ \mathrm{m}$. (b) On the same graph, plot the potential of a point charge with a 1-C charge at the origin. (c) Which potential is stronger near the rod? (d) What happens to the difference as the distance increases? Interpret your result.

Glossary

electric dipole

system of two equal but opposite charges a fixed distance apart

electric dipole moment

quantity defined as $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$ for all dipoles, where the vector points from the negative to positive charge

Determining Field from Potential

By the end of this section, you will be able to:

- Explain how to calculate the electric field in a system from the given potential
- Calculate the electric field in a given direction from a given potential
- Calculate the electric field throughout space from a given potential

Recall that we were able, in certain systems, to calculate the potential by integrating over the electric field. As you may already suspect, this means that we may calculate the electric field by taking derivatives of the potential, although going from a scalar to a vector quantity introduces some interesting wrinkles. We frequently need $\overrightarrow{\mathbf{E}}$ to calculate the force in a system; since it is often simpler to calculate the potential directly, there are systems in which it is useful to calculate *V* and then derive $\overrightarrow{\mathbf{E}}$ from it.

In general, regardless of whether the electric field is uniform, it points in the direction of decreasing potential, because the force on a positive charge is in the direction of $\overrightarrow{\mathbf{E}}$ and also in the direction of lower potential *V*. Furthermore, the magnitude of $\overrightarrow{\mathbf{E}}$ equals the rate of decrease of *V* with distance. The faster *V* decreases over distance, the greater the electric field. This gives us the following result.

Note:

Relationship between Voltage and Uniform Electric Field

In equation form, the relationship between voltage and uniform electric field is

Equation:

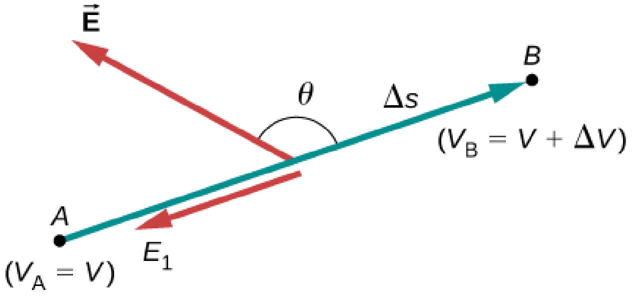
$$E=-\frac{\Delta V}{\Delta s}$$

where Δs is the distance over which the change in potential ΔV takes place. The minus sign tells us that *E* points in the direction of decreasing potential. The electric field is said to be the gradient (as in grade or slope) of the electric potential.

For continually changing potentials, ΔV and Δs become infinitesimals, and we need differential calculus to determine the electric field. As shown in [link], if we treat the distance Δs as very small so that the electric field is essentially constant over it, we find that

Equation:

$$E_s=-\frac{dV}{ds}.$$



The electric field component along the displacement Δs is given by $E = -\frac{\Delta V}{\Delta s}$. Note that A and B are assumed to be so close together that the field is constant along Δs.

Therefore, the electric field components in the Cartesian directions are given by

Note:
Equation:

$$E_x = -\frac{\partial V}{\partial x},\ E_y = -\frac{\partial V}{\partial y},\ E_z = -\frac{\partial V}{\partial z}.$$

This allows us to define the “grad” or “del” vector operator, which allows us to compute the gradient in one step. In Cartesian coordinates, it takes the form

Note:
Equation:

$$\vec{\nabla} = \mathbf{\hat{i}}\frac{\partial}{\partial x} + \mathbf{\hat{j}}\frac{\partial}{\partial y} + \mathbf{\hat{k}}\frac{\partial}{\partial z}.$$

With this notation, we can calculate the electric field from the potential with

Note:
Equation:

$$\overrightarrow{\mathbf{E}} = -\vec{\nabla}V,$$

a process we call calculating the gradient of the potential.

If we have a system with either cylindrical or spherical symmetry, we only need to use the del operator in the appropriate coordinates:

Note:
Equation:

$$\text{Cylindrical: } \vec{\nabla} = \hat{\mathbf{r}}\frac{\partial}{\partial r} + \widehat{\varphi}\frac{1}{r}\frac{\partial}{\partial \varphi} + \widehat{\mathbf{z}}\frac{\partial}{\partial z}$$

Note:
Equation:

$$\text{Spherical: } \vec{\nabla} = \hat{\mathbf{r}}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \widehat{\varphi}\frac{1}{r\sin\theta}\frac{\partial}{\partial \varphi}$$

Example:
Electric Field of a Point Charge
Calculate the electric field of a point charge from the potential.
Strategy
The potential is known to be $V = k\frac{q}{r}$, which has a spherical symmetry. Therefore, we use the spherical del operator in the formula $\overrightarrow{\mathbf{E}} = -\vec{\nabla}V$.
Solution
Performing this calculation gives us
Equation:

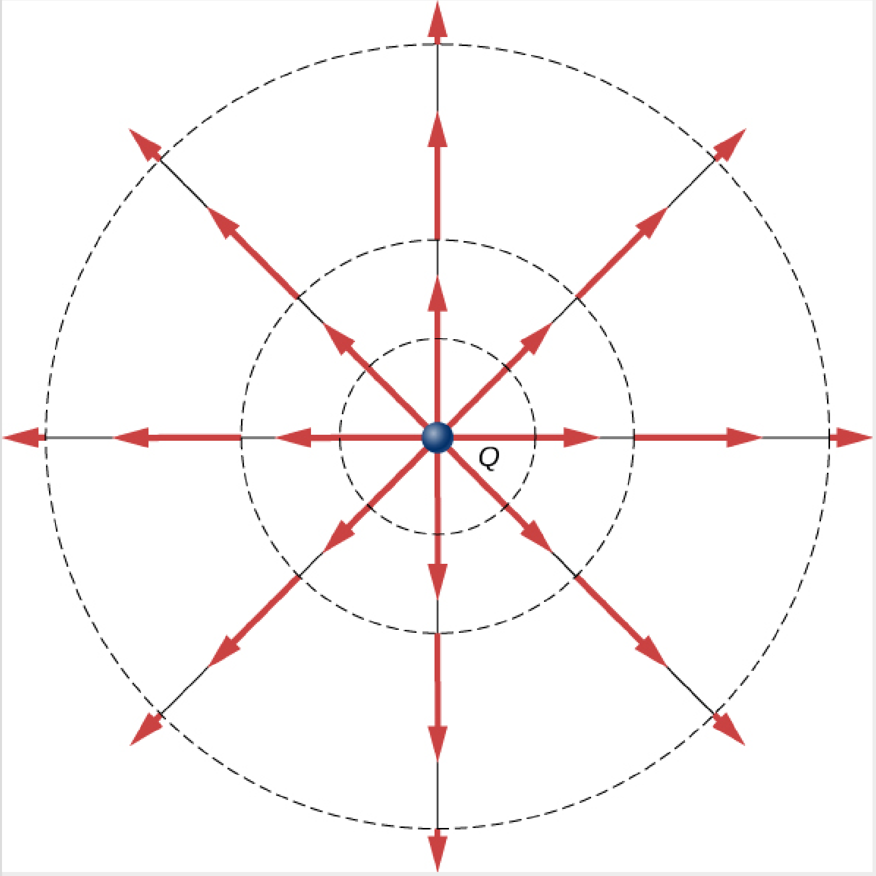
$$\overrightarrow{\mathbf{E}} = -\left(\hat{\mathbf{r}}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \widehat{\varphi}\frac{1}{r\sin\theta}\frac{\partial}{\partial \varphi}\right)k\frac{q}{r} = -kq\left(\hat{\mathbf{r}}\frac{\partial}{\partial r}\frac{1}{r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta}\frac{1}{r} + \widehat{\varphi}\frac{1}{r\sin\theta}\frac{\partial}{\partial \varphi}\frac{1}{r}\right).$$

This equation simplifies to
Equation:

$$\overrightarrow{\mathbf{E}} = -kq\left(\hat{\mathbf{r}}\frac{-1}{r^2} + \hat{\theta}0 + \widehat{\varphi}0\right) = k\frac{q}{r^2}\hat{\mathbf{r}}$$

as expected.
Significance

We not only obtained the equation for the electric field of a point particle that we’ve seen before, we also have a demonstration that $\vec{\mathbf{E}}$ points in the direction of decreasing potential, as shown in [link].

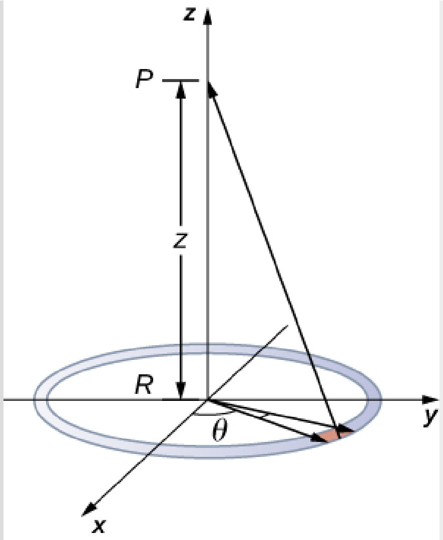


Electric field vectors inside and outside a uniformly charged sphere.

Example:

Electric Field of a Ring of Charge

Use the potential found in [link] to calculate the electric field along the axis of a ring of charge ([link]).



We want to calculate the electric field from the electric potential due to a ring charge.

Strategy

In this case, we are only interested in one dimension, the *z*-axis. Therefore, we use $E_z = -\frac{\partial V}{\partial z}$ with the potential $V = k\frac{q_{\text{tot}}}{\sqrt{z^2+R^2}}$ found previously.

Solution

Taking the derivative of the potential yields

Equation:

$$E_z = -\frac{\partial}{\partial z}\frac{kq_{\text{tot}}}{\sqrt{z^2+R^2}} = k\frac{q_{\text{tot}}z}{\left(z^2+R^2\right)^{3/2}}.$$

Significance

Again, this matches the equation for the electric field found previously. It also demonstrates a system in which using the full del operator is not necessary.

Note:

Exercise:

Problem:

Check Your Understanding Which coordinate system would you use to calculate the electric field of a dipole?

Solution:

Any, but cylindrical is closest to the symmetry of a dipole.

Summary

- Just as we may integrate over the electric field to calculate the potential, we may take the derivative of the potential to calculate the electric field.
- This may be done for individual components of the electric field, or we may calculate the entire electric field vector with the gradient operator.

Conceptual Questions

Exercise:

Problem:

If the electric field is zero throughout a region, must the electric potential also be zero in that region?

Solution:

No. It will be constant, but not necessarily zero.

Exercise:

Problem:

Explain why knowledge of $\vec{\mathbf{E}}(x, y, z)$ is not sufficient to determine *V*(*x*,*y*,*z*). What about the other way around?

Problems

Exercise:

Problem:

Throughout a region, equipotential surfaces are given by $z = \text{constant}$. The surfaces are equally spaced with $V = 100\text{ V}$ for $z = 0.00\text{ m}$, $V = 200\text{ V}$ for $z = 0.50\text{ m}$, $V = 300\text{ V}$ for $z = 1.00\text{ m}$. What is the electric field in this region?

Solution:

The problem is describing a uniform field, so $E = 200\text{ V/m}$ in the –*z*-direction.

Exercise:

Problem:

In a particular region, the electric potential is given by $V = -xy^2z + 4xy$. What is the electric field in this region?

Exercise:

Problem: Calculate the electric field of an infinite line charge, throughout space.

Solution:

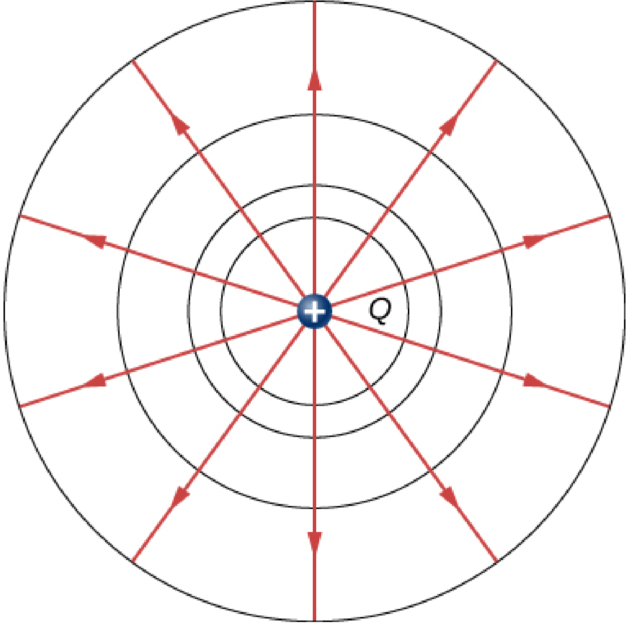
Apply $\overrightarrow{\mathbf{E}} = -\vec{\nabla} V$ with $\vec{\nabla} = \hat{\mathbf{r}} \frac{\partial}{\partial r} + \widehat{\varphi} \frac{1}{r} \frac{\partial}{\partial \varphi} + \widehat{\mathbf{z}} \frac{\partial}{\partial z}$ to the potential calculated earlier, $V = -2k\lambda \ln s :$ $\overrightarrow{\mathbf{E}} = 2k\lambda \frac{1}{r} \hat{\mathbf{r}}$ as expected.

Equipotential Surfaces and Conductors

By the end of this section, you will be able to:

- Define equipotential surfaces and equipotential lines
- Explain the relationship between equipotential lines and electric field lines
- Map equipotential lines for one or two point charges
- Describe the potential of a conductor
- Compare and contrast equipotential lines and elevation lines on topographic maps

We can represent electric potentials (voltages) pictorially, just as we drew pictures to illustrate electric fields. This is not surprising, since the two concepts are related. Consider [link], which shows an isolated positive point charge and its electric field lines, which radiate out from a positive charge and terminate on negative charges. We use red arrows to represent the magnitude and direction of the electric field, and we use black lines to represent places where the electric potential is constant. These are called **equipotential surfaces** in three dimensions, or **equipotential lines** in two dimensions. The term *equipotential* is also used as a noun, referring to an equipotential line or surface. The potential for a point charge is the same anywhere on an imaginary sphere of radius r surrounding the charge. This is true because the potential for a point charge is given by $V = kq/r$ and thus has the same value at any point that is a given distance r from the charge. An equipotential sphere is a circle in the two-dimensional view of [link]. Because the electric field lines point radially away from the charge, they are perpendicular to the equipotential lines.



An isolated point charge Q with its electric field lines in red and equipotential lines in black. The potential is the same along each equipotential line, meaning that no work is required to move a charge anywhere along one of those lines. Work is needed to move a charge from one is needed to move a charge from one equipotential line to another. Equipotential lines are perpendicular to electric field lines in every case. For a three-dimensional version, explore the first media link.

It is important to note that *equipotential lines are always perpendicular to electric field lines*. No work is required to move a charge along an equipotential, since $\Delta V = 0$. Thus, the work is

Equation:

$$W = -\Delta U = -q\Delta V = 0.$$

Work is zero if the direction of the force is perpendicular to the displacement. Force is in the same direction as *E*, so motion along an equipotential must be perpendicular to *E*. More precisely, work is related to the electric field by

Equation:

$$W = \overrightarrow{\mathbf{F}} \cdot \overrightarrow{\mathbf{d}} = q\overrightarrow{\mathbf{E}} \cdot \overrightarrow{\mathbf{d}} = qEd \cos\theta = 0.$$

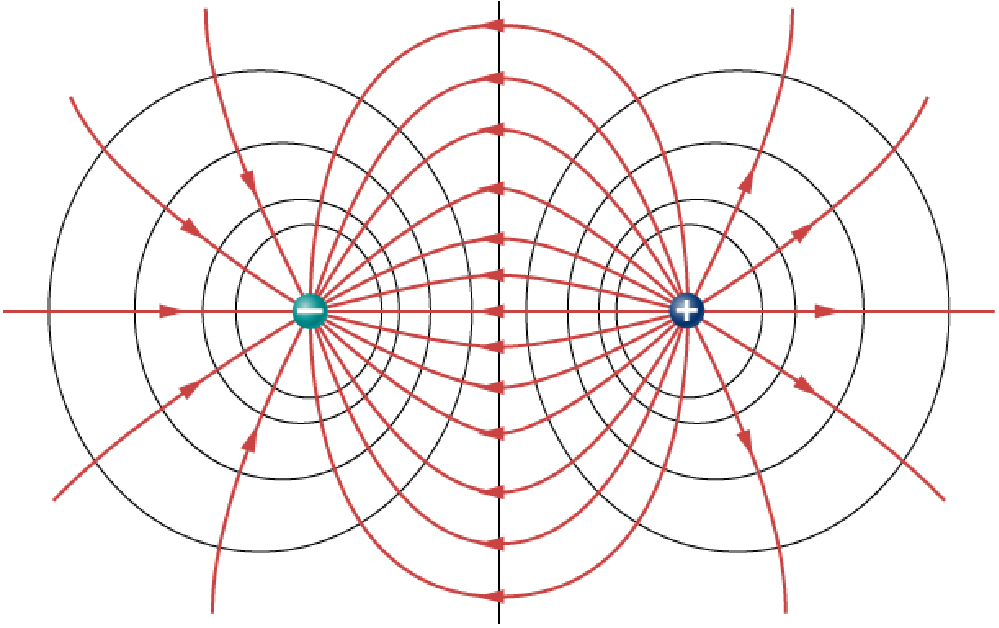
Note that in this equation, *E* and *F* symbolize the magnitudes of the electric field and force, respectively. Neither *q* nor *E* is zero; *d* is also not zero. So cos θ must be 0, meaning θ must be 90º. In other words, motion along an equipotential is perpendicular to *E*.

One of the rules for static electric fields and conductors is that the electric field must be perpendicular to the surface of any conductor. This implies that a *conductor is an equipotential surface in static situations*. There can be no voltage difference across the surface of a conductor, or charges will flow. One of the uses of this fact is that a conductor can be fixed at what we consider zero volts by connecting it to the earth with a good conductor—a process called **grounding**. Grounding can be a useful safety tool. For example, grounding the metal case of an electrical appliance ensures that it is at zero volts relative to Earth.

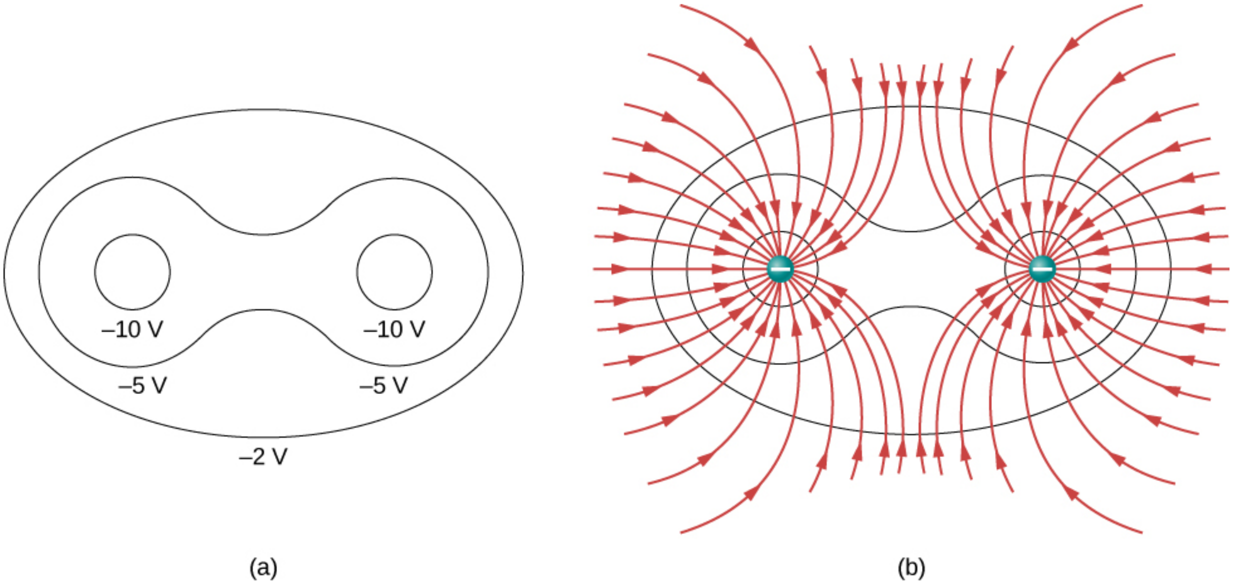
Because a conductor is an equipotential, it can replace any equipotential surface. For example, in [link], a charged spherical conductor can replace the point charge, and the electric field and potential surfaces outside of it will be unchanged, confirming the contention that a spherical charge distribution is equivalent to a point charge at its center.

[link] shows the electric field and equipotential lines for two equal and opposite charges. Given the electric field lines, the equipotential lines can be drawn simply by making them perpendicular to the electric field lines. Conversely, given the equipotential lines, as in [link]

(a), the electric field lines can be drawn by making them perpendicular to the equipotentials, as in [link](b).

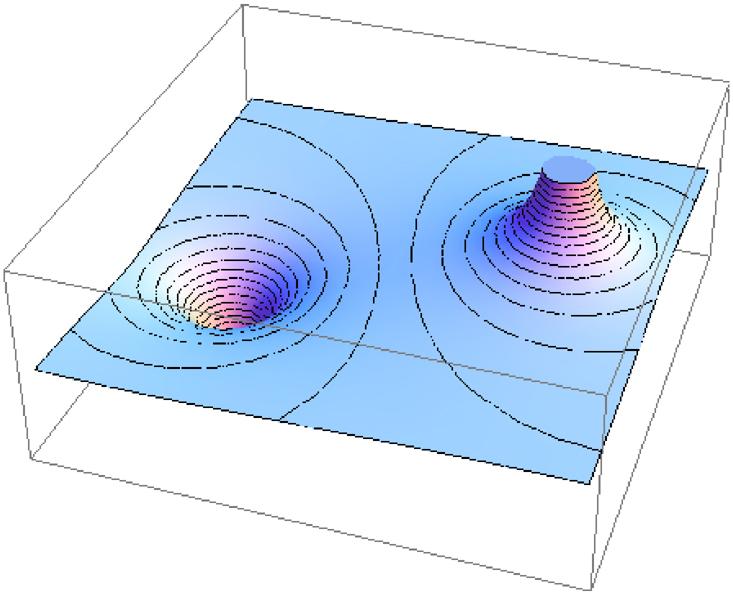


The electric field lines and equipotential lines for two equal but opposite charges. The equipotential lines can be drawn by making them perpendicular to the electric field lines, if those are known. Note that the potential is greatest (most positive) near the positive charge and least (most negative) near the negative charge. For a three-dimensional version, explore the first media link.



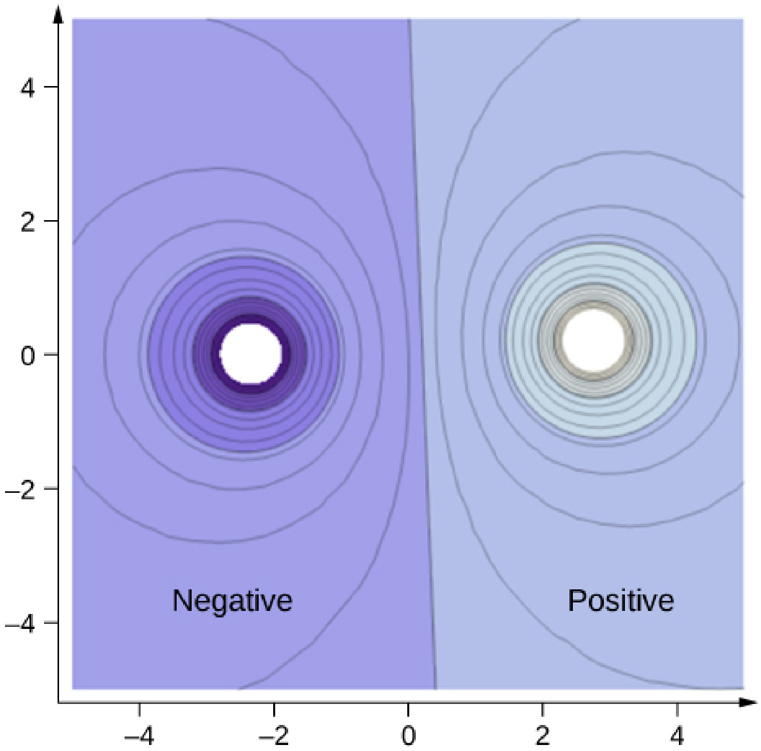
(a) These equipotential lines might be measured with a voltmeter in a laboratory experiment. (b) The corresponding electric field lines are found by drawing them perpendicular to the equipotentials. Note that these fields are consistent with two equal negative charges. For a three-dimensional version, play with the first media link.

To improve your intuition, we show a three-dimensional variant of the potential in a system with two opposing charges. [link] displays a three-dimensional map of electric potential, where lines on the map are for equipotential surfaces. The hill is at the positive charge, and the trough is at the negative charge. The potential is zero far away from the charges. Note that the cut off at a particular potential implies that the charges are on conducting spheres with a finite radius.



Electric potential map of two opposite charges of equal magnitude on conducting spheres. The potential is negative near the negative charge and positive near the positive charge.

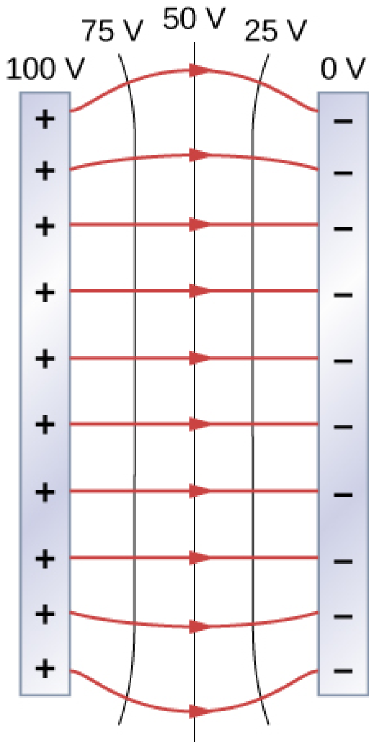
A two-dimensional map of the cross-sectional plane that contains both charges is shown in [link]. The line that is equidistant from the two opposite charges corresponds to zero potential, since at the points on the line, the positive potential from the positive charge cancels the negative potential from the negative charge. Equipotential lines in the cross-sectional plane are closed loops, which are not necessarily circles, since at each point, the net potential is the sum of the potentials from each charge.



A cross-section of the electric potential map of two opposite charges of equal magnitude. The potential is negative near the negative charge and positive near the positive charge.

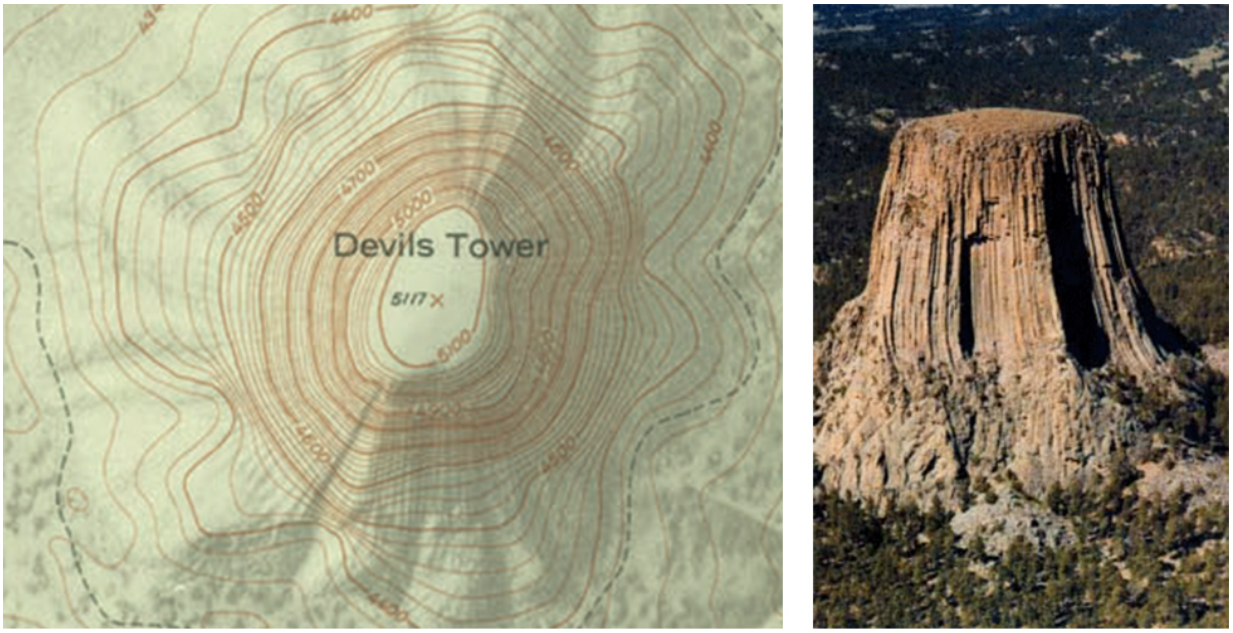
Note:
View this simulation to observe and modify the equipotential surfaces and electric fields for many standard charge configurations. There’s a lot to explore.

One of the most important cases is that of the familiar parallel conducting plates shown in [link]. Between the plates, the equipotentials are evenly spaced and parallel. The same field could be maintained by placing conducting plates at the equipotential lines at the potentials shown.



The electric field and equipotential lines between two metal plates. Note that the electric field is perpendicular to the equipotentials and hence normal to the plates at their surface as well as in the center of the region between them.

Consider the parallel plates in [link]. These have equipotential lines that are parallel to the plates in the space between and evenly spaced. An example of this (with sample values) is given in [link]. We could draw a similar set of equipotential isolines for gravity on the hill shown in [link]. If the hill has any extent at the same slope, the isolines along that extent would be parallel to each other. Furthermore, in regions of constant slope, the isolines would be evenly spaced. An example of real topographic lines is shown in [link].



(a) (b)

A topographical map along a ridge has roughly parallel elevation lines, similar to the equipotential lines in [link]. (a) A topographical map of Devil’s Tower, Wyoming. Lines that are close together indicate very steep terrain. (b) A perspective photo of Devil’s Tower shows just how steep its sides are. Notice the top of the tower has the same shape as the center of the topographical map.

Example:

Calculating Equipotential Lines

You have seen the equipotential lines of a point charge in [link]. How do we calculate them? For example, if we have a $+10\text{-nC}$ charge at the origin, what are the equipotential surfaces at which the potential is (a) 100 V, (b) 50 V, (c) 20 V, and (d) 10 V?

Strategy

Set the equation for the potential of a point charge equal to a constant and solve for the remaining variable(s). Then calculate values as needed.

Solution

In $V=k\frac{q}{r}$, let V be a constant. The only remaining variable is r; hence, $r=k\frac{q}{V}=\text{constant}$. Thus, the equipotential surfaces are spheres about the origin. Their locations are:

a. $r=k\frac{q}{V}=\left(8.99\,\times\,10^{9}\,\text{Nm}^{2}/\text{C}^{2}\right)\frac{(10\times10^{-9}\,\text{C})}{100\text{ V}}=0.90\text{ m};$

b. $r=k\frac{q}{V}=\left(8.99\,\times\,10^{9}\,\text{Nm}^{2}/\text{C}^{2}\right)\frac{(10\times10^{-9}\,\text{C})}{50\text{ V}}=1.8\text{ m};$

c. $r = k\frac{q}{V} = \left(8.99 \times 10^9 \,\mathrm{Nm^2/C^2}\right)\frac{\left(10 \times 10^{-9}\,\mathrm{C}\right)}{20\,\mathrm{V}} = 4.5\,\mathrm{m};$

d. $r = k\frac{q}{V} = \left(8.99 \times 10^9 \,\mathrm{Nm^2/C^2}\right)\frac{\left(10 \times 10^{-9}\,\mathrm{C}\right)}{10\,\mathrm{V}} = 9.0\,\mathrm{m}.$

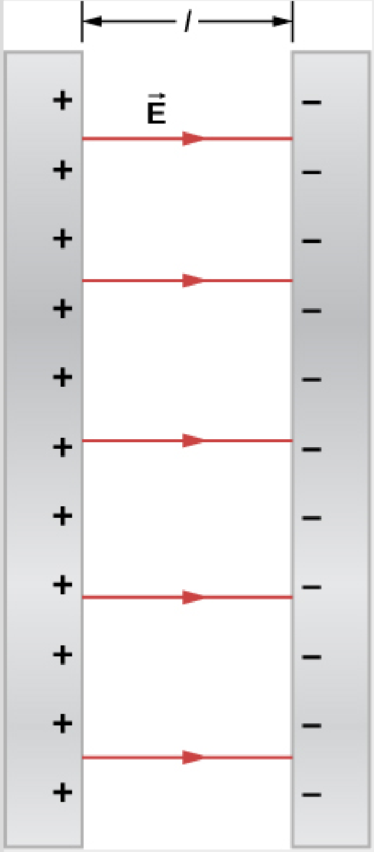
Significance

This means that equipotential surfaces around a point charge are spheres of constant radius, as shown earlier, with well-defined locations.

Example:

Potential Difference between Oppositely Charged Parallel Plates

Two large conducting plates carry equal and opposite charges, with a surface charge density σ of magnitude $6.81 \times 10^{-7}\,\mathrm{C/m^2}$, as shown in [link]. The separation between the plates is $l = 6.50\,\mathrm{mm}$. (a) What is the electric field between the plates? (b) What is the potential difference between the plates? (c) What is the distance between equipotential planes which differ by 100 V?



The electric field between oppositely charged parallel plates.

A portion is released
at the positive plate.

Strategy
(a) Since the plates are described as “large” and the distance between them is not, we will approximate each of them as an infinite plane, and apply the result from Gauss’s law in the previous chapter.

(b) Use $\Delta V_{AB} = -\int_A^B \overrightarrow{\mathbf{E}} \cdot d\vec{\mathbf{l}}$.

(c) Since the electric field is constant, find the ratio of 100 V to the total potential difference; then calculate this fraction of the distance.

Solution

a. The electric field is directed from the positive to the negative plate as shown in the figure, and its magnitude is given by
Equation:

$$E = \frac{\sigma}{\varepsilon_0} = \frac{6.81 \times 10^{-7}\,\mathrm{C/m^2}}{8.85 \times 10^{-12}\,\mathrm{C^2/N\cdot m^2}} = 7.69 \times 10^4\,\mathrm{V/m}.$$

b. To find the potential difference ΔV between the plates, we use a path from the negative to the positive plate that is directed against the field. The displacement vector $d\vec{\mathbf{l}}$ and the electric field $\vec{\mathbf{E}}$ are antiparallel so $\vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = -E\,dl$. The potential difference between the positive plate and the negative plate is then
Equation:

$$\Delta V = -\int E \cdot dl = E\int dl = El = (7.69 \times 10^4\,\mathrm{V/m})(6.50 \times 10^{-3}\,\mathrm{m}) = 500\,\mathrm{V}.$$

c. The total potential difference is 500 V, so 1/5 of the distance between the plates will be the distance between 100-V potential differences. The distance between the plates is 6.5 mm, so there will be 1.3 mm between 100-V potential differences.

Significance
You have now seen a numerical calculation of the locations of equipotentials between two charged parallel plates.

Note:
Exercise:

Problem:

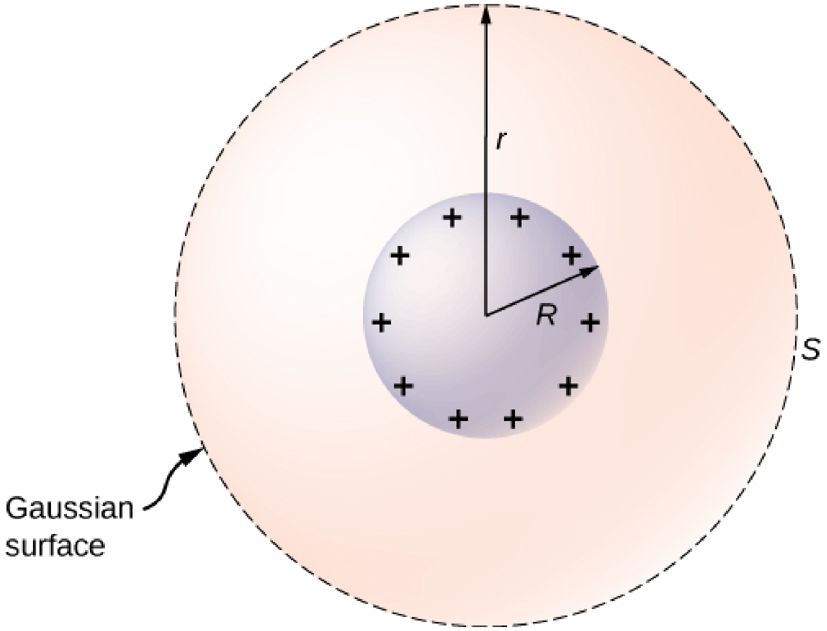
Check Your Understanding What are the equipotential surfaces for an infinite line charge?

Solution:

infinite cylinders of constant radius, with the line charge as the axis

Distribution of Charges on Conductors

In [link] with a point charge, we found that the equipotential surfaces were in the form of spheres, with the point charge at the center. Given that a conducting sphere in electrostatic equilibrium is a spherical equipotential surface, we should expect that we could replace one of the surfaces in [link] with a conducting sphere and have an identical solution outside the sphere. Inside will be rather different, however.



An isolated conducting sphere.

To investigate this, consider the isolated conducting sphere of [link] that has a radius *R* and an excess charge *q*. To find the electric field both inside and outside the sphere, note that the sphere is isolated, so its surface change distribution and the electric field of that distribution

are spherically symmetric. We can therefore represent the field as $\vec{\mathbf{E}} = E(r)\hat{\mathbf{r}}$. To calculate *E(r)*, we apply Gauss's law over a closed spherical surface *S* of radius *r* that is concentric with the conducting sphere. Since *r* is constant and $\widehat{\mathbf{n}} = \hat{\mathbf{r}}$ on the sphere,
Equation:

$$\oint_S \vec{\mathbf{E}} \cdot \widehat{\mathbf{n}}\, da = E(r) \oint da = E(r)\, 4\pi r^2.$$

For $r \ < \ R$, *S* is within the conductor, so recall from our previous study of Gauss's law that $q_{\text{enc}} = 0$ and Gauss's law gives $E\,(r) = 0$, as expected inside a conductor at equilibrium. If $r \ > \ R$, *S* encloses the conductor so $q_{\text{enc}} = q$. From Gauss's law,
Equation:

$$E(r)\, 4\pi r^2 = \frac{q}{\varepsilon_0}.$$

The electric field of the sphere may therefore be written as
Equation:

$$\begin{array}{lllr} E & = & 0 & (r < R), \\ E & = & \frac{1}{4\pi\varepsilon_0}\,\frac{q}{r^2}\hat{\mathbf{r}} & (r \geq R). \end{array}$$

As expected, in the region $r \geq R$, the electric field due to a charge *q* placed on an isolated conducting sphere of radius *R* is identical to the electric field of a point charge *q* located at the center of the sphere.

To find the electric potential inside and outside the sphere, note that for $r \geq R$, the potential must be the same as that of an isolated point charge *q* located at $r = 0$,
Equation:

$$V(r) = \frac{1}{4\pi\varepsilon_0}\,\frac{q}{r}\,(r \geq R)$$

simply due to the similarity of the electric field.

For $r < R, E = 0$, so *V*(*r*) is constant in this region. Since $V(R) = q/4\pi\varepsilon_0 R$,
Equation:

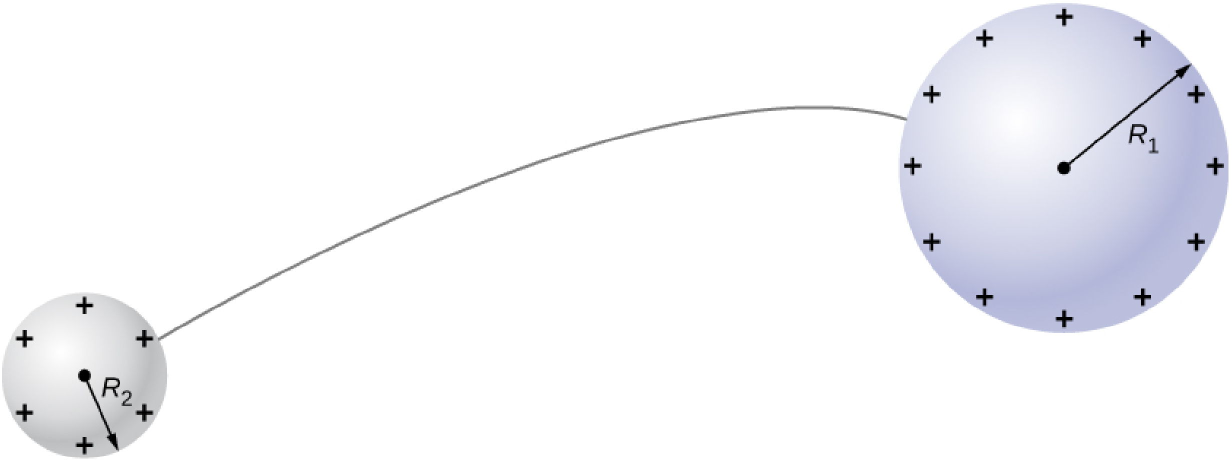
$$V(r) = \frac{1}{4\pi\varepsilon_0}\,\frac{q}{R}\,(r < R).$$

We will use this result to show that

Equation:

$$\sigma_1 R_1 = \sigma_2 R_2,$$

for two conducting spheres of radii R_1 and R_2, with surface charge densities σ_1 and σ_2 respectively, that are connected by a thin wire, as shown in [link]. The spheres are sufficiently separated so that each can be treated as if it were isolated (aside from the wire). Note that the connection by the wire means that this entire system must be an equipotential.



Two conducting spheres are connected by a thin conducting wire.

We have just seen that the electrical potential at the surface of an isolated, charged conducting sphere of radius R is

Equation:

$$V = \frac{1}{4\pi\varepsilon_0}\frac{q}{R}.$$

Now, the spheres are connected by a conductor and are therefore at the same potential; hence

Equation:

$$\frac{1}{4\pi\varepsilon_0}\frac{q_1}{R_1} = \frac{1}{4\pi\varepsilon_0}\frac{q_2}{R_2},$$

and

Equation:

$$\frac{q_1}{R_1} = \frac{q_2}{R_2}.$$

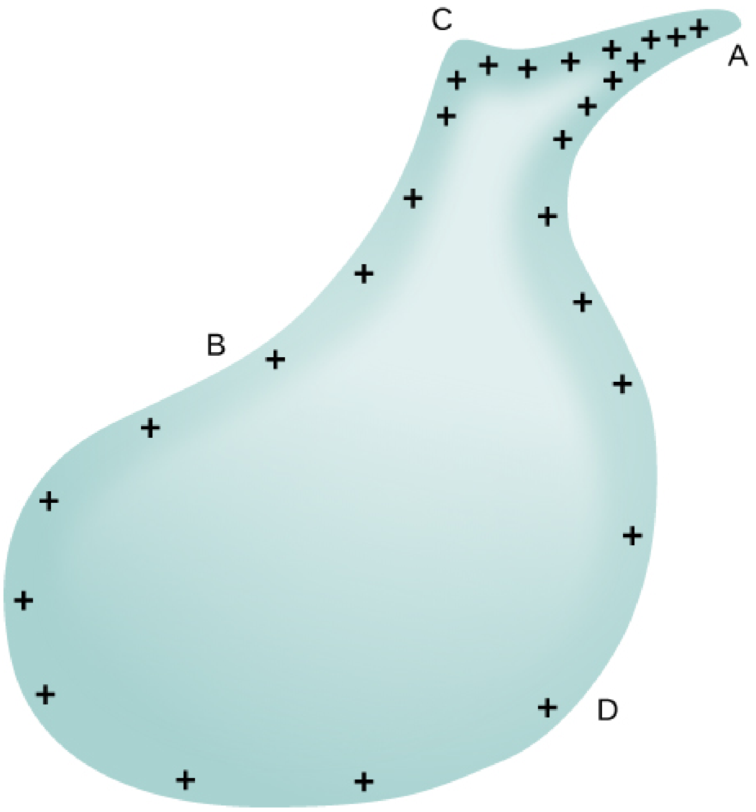
The net charge on a conducting sphere and its surface charge density are related by $q = \sigma(4\pi R^2)$. Substituting this equation into the previous one, we find

Equation:

$$\sigma_1 R_1 = \sigma_2 R_2.$$

Obviously, two spheres connected by a thin wire do not constitute a typical conductor with a variable radius of curvature. Nevertheless, this result does at least provide a qualitative idea of how charge density varies over the surface of a conductor. The equation indicates that where the radius of curvature is large (points *B* and *D* in [link]), σ and E are small.

Similarly, the charges tend to be denser where the curvature of the surface is greater, as demonstrated by the charge distribution on oddly shaped metal ([link]). The surface charge density is higher at locations with a small radius of curvature than at locations with a large radius of curvature.



The surface charge density and the electric field of a conductor are greater at regions with smaller radii of curvature.

A practical application of this phenomenon is the lightning rod, which is simply a grounded metal rod with a sharp end pointing upward. As positive charge accumulates in the ground due to a negatively charged cloud overhead, the electric field around the sharp point gets very large. When the field reaches a value of approximately $3.0 \times 10^6 \mathrm{~N/C}$ (the *dielectric strength* of the air), the free ions in the air are accelerated to such high energies that their collisions with air molecules actually ionize the molecules. The resulting free electrons in the air then flow through the rod to Earth, thereby neutralizing some of the positive charge. This keeps the electric field between the cloud and the ground from getting large enough to produce a lightning bolt in the region around the rod.

An important application of electric fields and equipotential lines involves the heart. The heart relies on electrical signals to maintain its rhythm. The movement of electrical signals causes the chambers of the heart to contract and relax. When a person has a heart attack, the movement of these electrical signals may be disturbed. An artificial pacemaker and a defibrillator can be used to initiate the rhythm of electrical signals. The equipotential lines around the heart, the thoracic region, and the axis of the heart are useful ways of monitoring the structure and functions of the heart. An electrocardiogram (ECG) measures the small electric signals being generated during the activity of the heart.

Note:
Play around with this simulation to move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more.

Summary

- An equipotential surface is the collection of points in space that are all at the same potential. Equipotential lines are the two-dimensional representation of equipotential surfaces.
- Equipotential surfaces are always perpendicular to electric field lines.
- Conductors in static equilibrium are equipotential surfaces.
- Topographic maps may be thought of as showing gravitational equipotential lines.

Conceptual Questions

Exercise:

Problem:

If two points are at the same potential, are there any electric field lines connecting them?

Solution:

no

Exercise:

Problem:

Suppose you have a map of equipotential surfaces spaced 1.0 V apart. What do the distances between the surfaces in a particular region tell you about the strength of the $\vec{\mathbf{E}}$ in that region?

Exercise:

Problem: Is the electric potential necessarily constant over the surface of a conductor?

Solution:

No; it might not be at electrostatic equilibrium.

Exercise:

Problem:

Under electrostatic conditions, the excess charge on a conductor resides on its surface. Does this mean that all of the conduction electrons in a conductor are on the surface?

Exercise:

Problem: Can a positively charged conductor be at a negative potential? Explain.

Solution:

Yes. It depends on where the zero reference for potential is. (Though this might be unusual.)

Exercise:

Problem: Can equipotential surfaces intersect?

Problems

Exercise:

Problem:

Two very large metal plates are placed 2.0 cm apart, with a potential difference of 12 V between them. Consider one plate to be at 12 V, and the other at 0 V. (a) Sketch the equipotential surfaces for 0, 4, 8, and 12 V. (b) Next sketch in some electric field lines, and confirm that they are perpendicular to the equipotential lines.

Exercise:

Problem:

A very large sheet of insulating material has had an excess of electrons placed on it to a surface charge density of $-3.00\ \mathrm{nC/m^2}$. (a) As the distance from the sheet increases, does the potential increase or decrease? Can you explain why without any calculations? Does the location of your reference point matter? (b) What is the shape of the equipotential surfaces? (c) What is the spacing between surfaces that differ by 1.00 V?

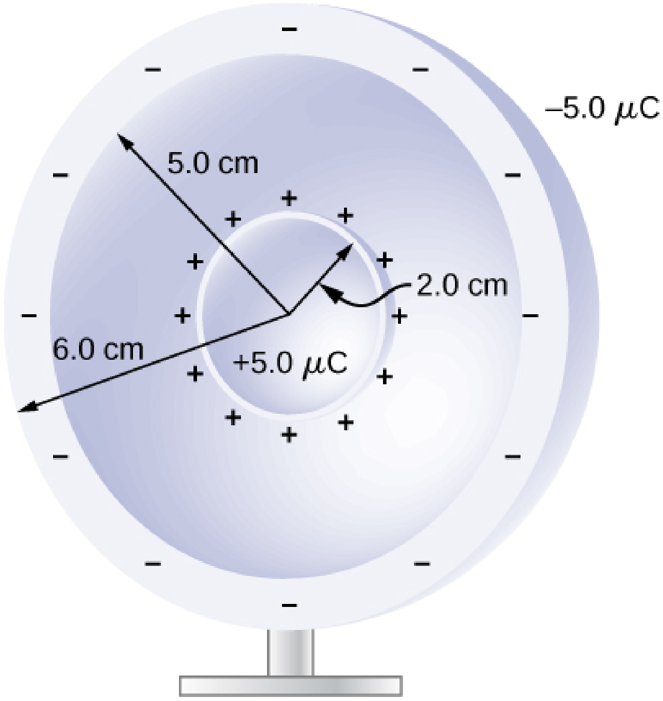
Solution:

a. increases; the constant (negative) electric field has this effect, the reference point only matters for magnitude; b. they are planes parallel to the sheet; c. 0.006 m/V

Exercise:

Problem:

A metallic sphere of radius 2.0 cm is charged with $+5.0\text{-}\mu\mathrm{C}$ charge, which spreads on the surface of the sphere uniformly. The metallic sphere stands on an insulated stand and is surrounded by a larger metallic spherical shell, of inner radius 5.0 cm and outer radius 6.0 cm. Now, a charge of $-5.0\text{-}\mu\mathrm{C}$ is placed on the inside of the spherical shell, which spreads out uniformly on the inside surface of the shell. If potential is zero at infinity, what is the potential of (a) the spherical shell, (b) the sphere, (c) the space between the two, (d) inside the sphere, and (e) outside the shell?



Exercise:

Problem:

Two large charged plates of charge density $\pm 30\ \mu\text{C/m}^2$ face each other at a separation of 5.0 mm. (a) Find the electric potential everywhere. (b) An electron is released from rest at the negative plate; with what speed will it strike the positive plate?

Solution:

a. from the previous chapter, the electric field has magnitude $\frac{\sigma}{\varepsilon_0}$ in the region between the plates and zero outside; defining the negatively charged plate to be at the origin and zero potential, with the positively charged plate located at $+5\ \text{mm}$ in the *z*-direction, $V = 1.7\ \times\ 10^4\ \text{V}$ so the potential is 0 for $z < 0,\ 1.7\ \times\ 10^4\ \text{V}\left(\frac{z}{5\ \text{mm}}\right)$ for $0 \leq z \leq 5\ \text{mm}$, $1.7\ \times\ 10^4\ \text{V}$ for $z > 5\ \text{mm}$;
b. $qV = \frac{1}{2}mv^2 \rightarrow v = 7.7\ \times\ 10^7\ \text{m/s}$

Exercise:

Problem:

A long cylinder of aluminum of radius *R* meters is charged so that it has a uniform charge per unit length on its surface of λ.

(a) Find the electric field inside and outside the cylinder. (b) Find the electric potential inside and outside the cylinder. (c) Plot electric field and electric potential as a function of distance from the center of the rod.

Exercise:

Problem:

Two parallel plates 10 cm on a side are given equal and opposite charges of magnitude $5.0 \times 10^{-9}\,\text{C}$. The plates are 1.5 mm apart. What is the potential difference between the plates?

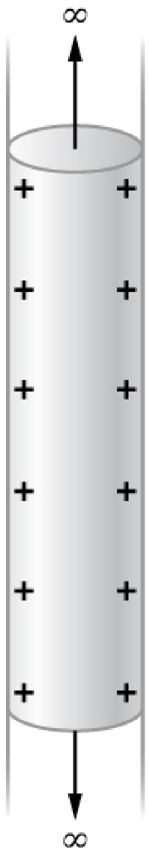
Solution:

$V = 85\,\text{V}$

Exercise:

Problem:

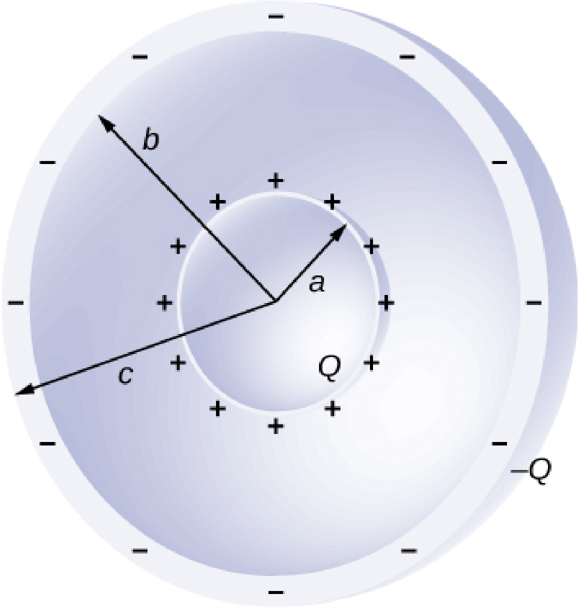
The surface charge density on a long straight metallic pipe is σ. What is the electric potential outside and inside the pipe? Assume the pipe has a diameter of $2a$.



Exercise:

Problem:

Concentric conducting spherical shells carry charges Q and –Q, respectively. The inner shell has negligible thickness. What is the potential difference between the shells?



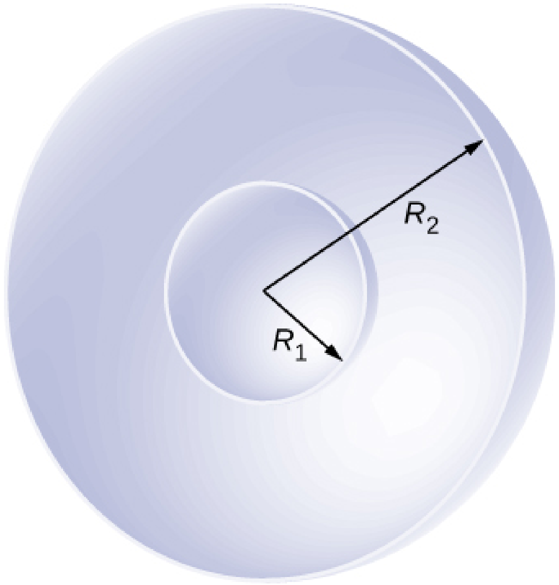
Solution:

In the region $a \leq r \leq b$, $\vec{\mathbf{E}} = \frac{kQ}{r^2}\hat{\mathbf{r}}$, and E is zero elsewhere; hence, the potential difference is $V = kQ\left(\frac{1}{a} - \frac{1}{b}\right)$.

Exercise:

Problem:

Shown below are two concentric spherical shells of negligible thicknesses and radii R_1 and R_2. The inner and outer shell carry net charges q_1 and q_2, respectively, where both q_1 and q_2 are positive. What is the electric potential in the regions (a) $r < R_1$, (b) $R_1 < r < R_2$, and (c) $r > R_2$?



Exercise:

Problem:

A solid cylindrical conductor of radius *a* is surrounded by a concentric cylindrical shell of inner radius *b*. The solid cylinder and the shell carry charges *Q* and –*Q*, respectively. Assuming that the length *L* of both conductors is much greater than *a* or *b*, what is the potential difference between the two conductors?

Solution:

From previous results $V_P - V_R = -2k\lambda \ln \frac{s_P}{s_R}$., note that *b* is a very convenient location to define the zero level of potential: $\Delta V = -2k\frac{Q}{L}\ln\frac{a}{b}$.

Glossary

equipotential line	two-dimensional representation of an equipotential surface
equipotential surface	surface (usually in three dimensions) on which all points are at the same potential
grounding	process of attaching a conductor to the earth to ensure that there is no potential difference between it and Earth

Applications of Electrostatics

By the end of this section, you will be able to:

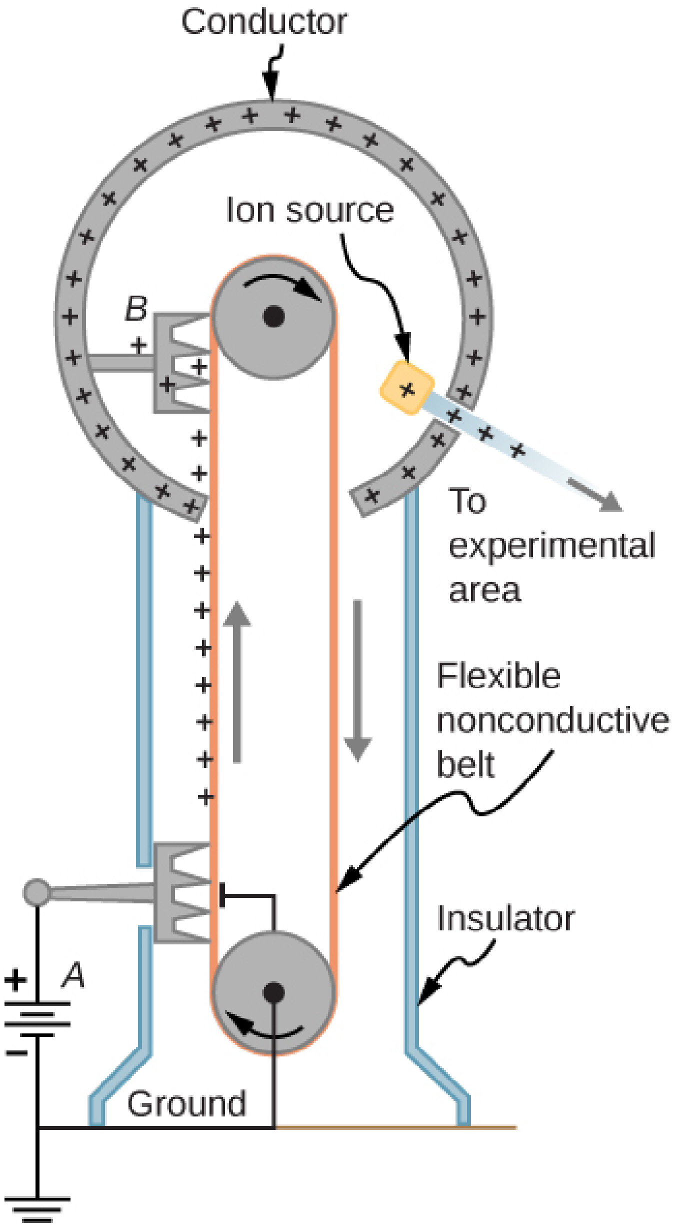
- Describe some of the many practical applications of electrostatics, including several printing technologies
- Relate these applications to Newton’s second law and the electric force

The study of electrostatics has proven useful in many areas. This module covers just a few of the many applications of electrostatics.

The Van de Graaff Generator

Van de Graaff generators (or Van de Graaffs) are not only spectacular devices used to demonstrate high voltage due to static electricity—they are also used for serious research. The first was built by Robert Van de Graaff in 1931 (based on original suggestions by Lord Kelvin) for use in nuclear physics research. [link] shows a schematic of a large research version. Van de Graaffs use both smooth and pointed surfaces, and conductors and insulators to generate large static charges and, hence, large voltages.

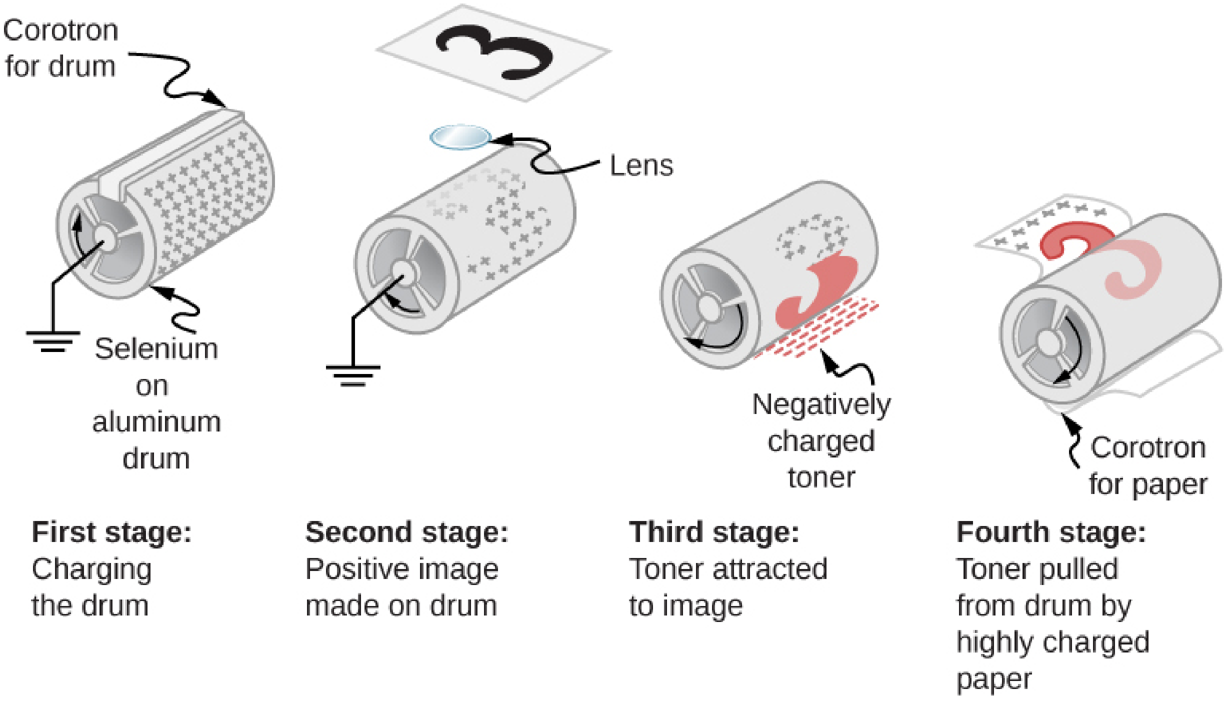
A very large excess charge can be deposited on the sphere because it moves quickly to the outer surface. Practical limits arise because the large electric fields polarize and eventually ionize surrounding materials, creating free charges that neutralize excess charge or allow it to escape. Nevertheless, voltages of 15 million volts are well within practical limits.


Conductor
Ion source
B
To experimental area
Flexible nonconductive belt
Insulator
A
Ground

Schematic of Van de Graaff generator. A battery (*A*) supplies excess positive charge to a pointed conductor, the points of which spray the charge onto a moving insulating belt near the bottom. The pointed conductor (*B*) on top in the large sphere picks up the charge. (The induced electric field at the points is so large that it removes the charge from the belt.) This can be done because the charge does not remain inside the conducting sphere but moves to its outside surface. An ion source inside the sphere produces positive ions, which are accelerated away from the positive sphere to high velocities.

Xerography

Most copy machines use an electrostatic process called **xerography**—a word coined from the Greek words *xeros* for dry and *graphos* for writing. The heart of the process is shown in simplified form in [link].


Corotron for drum
Selenium on aluminum drum
Lens
Negatively charged toner
Corotron for paper
First stage: Charging the drum
Second stage: Positive image made on drum
Third stage: Toner attracted to image
Fourth stage: Toner pulled from drum by highly charged paper

Xerography is a dry copying process based on electrostatics. The major steps in the process are the charging of the photoconducting drum, transfer of an image, creating a positive charge duplicate, attraction of toner to the charged parts of the drum, and transfer of toner to the paper. Not shown are heat treatment of the paper and cleansing of the drum for the next copy.

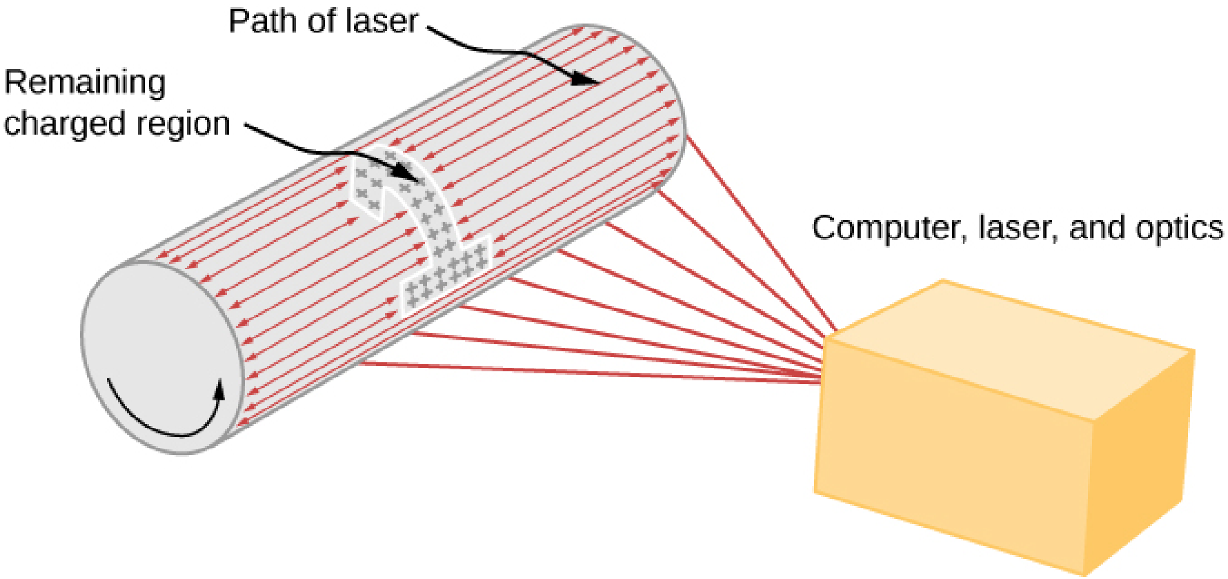
A selenium-coated aluminum drum is sprayed with positive charge from points on a device called a corotron. Selenium is a substance with an interesting property—it is a **photoconductor**. That is, selenium is an insulator when in the dark and a conductor when exposed to light.

In the first stage of the xerography process, the conducting aluminum drum is grounded so that a negative charge is induced under the thin layer of uniformly positively charged selenium. In the second stage, the surface of the drum is exposed to the image of whatever is to be copied. In locations where the image is light, the selenium becomes conducting, and the positive charge is neutralized. In dark areas, the positive charge remains, so the image has been transferred to the drum.

The third stage takes a dry black powder, called toner, and sprays it with a negative charge so that it is attracted to the positive regions of the drum. Next, a blank piece of paper is given a greater positive charge than on the drum so that it will pull the toner from the drum. Finally, the paper and electrostatically held toner are passed through heated pressure rollers, which melt and permanently adhere the toner to the fibers of the paper.

Laser Printers

Laser printers use the xerographic process to make high-quality images on paper, employing a laser to produce an image on the photoconducting drum as shown in [link]. In its most common application, the laser printer receives output from a computer, and it can achieve high-quality output because of the precision with which laser light can be controlled. Many laser printers do significant information processing, such as making sophisticated letters or fonts, and in the past may have contained a computer more powerful than the one giving them the raw data to be printed.

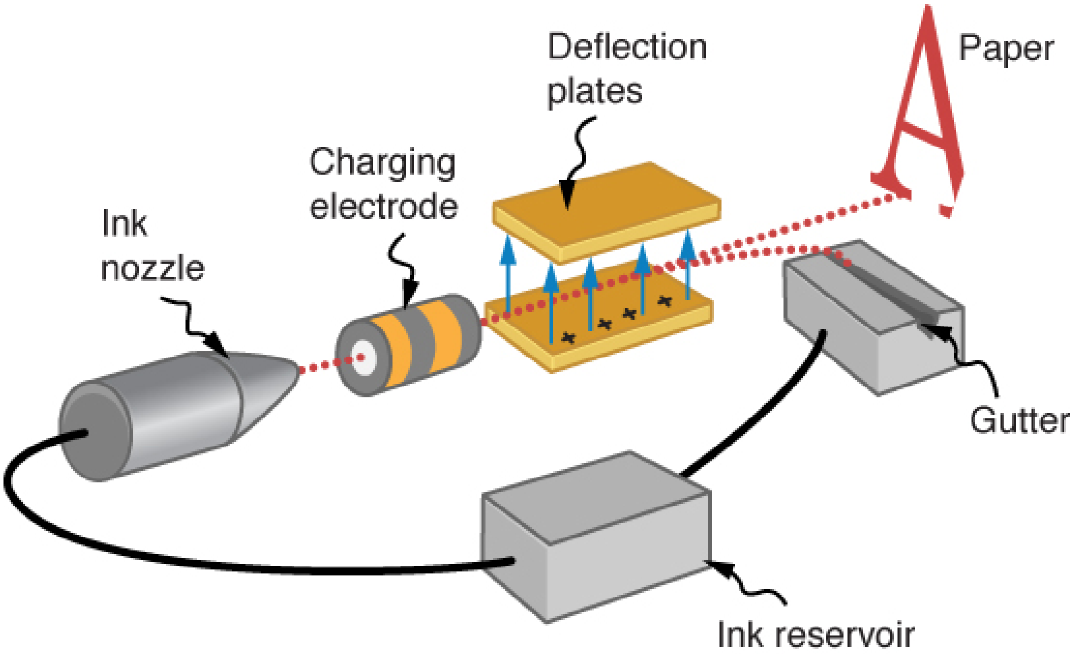


In a laser printer, a laser beam is scanned across a photoconducting drum, leaving a positively charged image. The other steps for charging the drum and transferring the image to paper are the same as in xerography. Laser light can be very precisely controlled, enabling laser printers to produce high-quality images.

Ink Jet Printers and Electrostatic Painting

The **ink jet printer**, commonly used to print computer-generated text and graphics, also employs electrostatics. A nozzle makes a fine spray of tiny ink droplets, which are then given an electrostatic charge ([link]).

Once charged, the droplets can be directed, using pairs of charged plates, with great precision to form letters and images on paper. Ink jet printers can produce color images by using a black jet and three other jets with primary colors, usually cyan, magenta, and yellow, much as a color television produces color. (This is more difficult with xerography, requiring multiple drums and toners.)



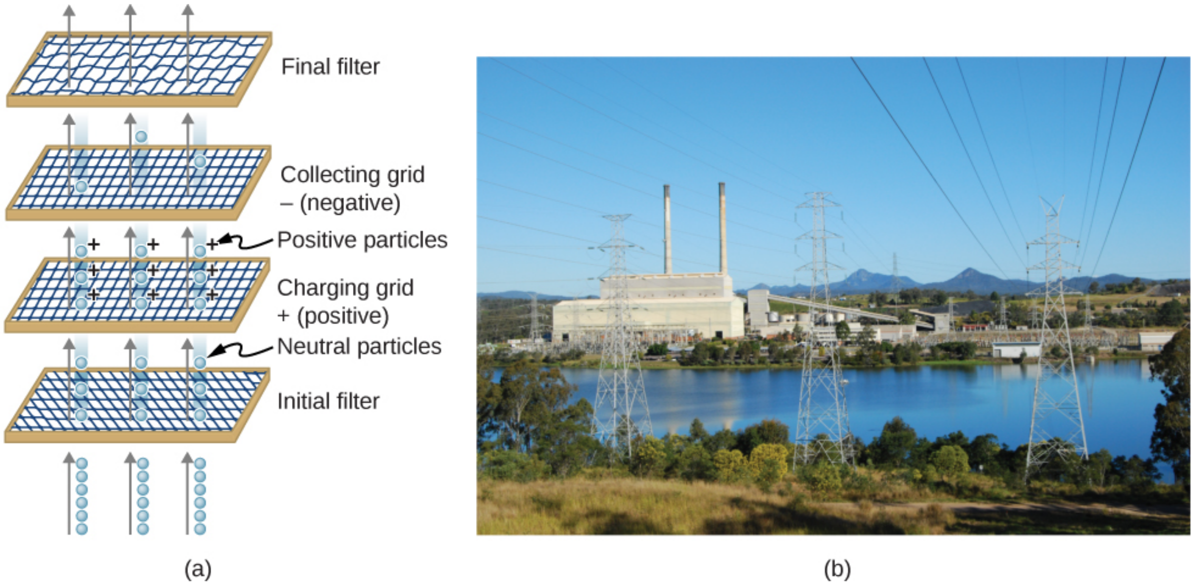
The nozzle of an ink-jet printer produces small ink droplets, which are sprayed with electrostatic charge. Various computer-driven devices are then used to direct the droplets to the correct positions on a page.

Electrostatic painting employs electrostatic charge to spray paint onto oddly shaped surfaces. Mutual repulsion of like charges causes the paint to fly away from its source. Surface tension forms drops, which are then attracted by unlike charges to the surface to be painted. Electrostatic painting can reach hard-to-get-to places, applying an even coat in a controlled manner. If the object is a conductor, the electric field is perpendicular to the surface, tending to bring the drops in perpendicularly. Corners and points on conductors will receive extra paint. Felt can similarly be applied.

Smoke Precipitators and Electrostatic Air Cleaning

Another important application of electrostatics is found in air cleaners, both large and small. The electrostatic part of the process places excess (usually positive) charge on smoke, dust, pollen, and other particles in the air and then passes the air through an oppositely charged grid that attracts and retains the charged particles ([link])

Large **electrostatic precipitators** are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.


Final filter
Collecting grid
– (negative)
Positive particles
Charging grid
+ (positive)
Neutral particles
Initial filter
(a)
(b)

(a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of electrostatic precipitators is seen by the absence of smoke from this power plant. (credit b: modification of work by “Cmdalgleish”/Wikimedia Commons)

Summary

- Electrostatics is the study of electric fields in static equilibrium.
- In addition to research using equipment such as a Van de Graaff generator, many practical applications of electrostatics exist, including

photocopiers, laser printers, ink jet printers, and electrostatic air filters.

Key Equations

Potential energy of a two-charge system	$U(r) = k\frac{qQ}{r}$
Work done to assemble a system of charges	$W_{12\cdots N} = \frac{k}{2}\sum_{i}^{N}\sum_{j}^{N}\frac{q_i q_j}{r_{ij}}$ for $i \neq j$
Potential difference	$\Delta V = \frac{\Delta U}{q}$ or $\Delta U = q\Delta V$
Electric potential	$V = \frac{U}{q} = -\int_{R}^{P}\vec{\mathbf{E}} \cdot d\vec{\mathbf{l}}$
Potential difference between two points	$\Delta V_{BA} = V_B - V_A = -\int_{A}^{B}\vec{\mathbf{E}} \cdot d\vec{\mathbf{l}}$
Electric potential of a point charge	$V = \frac{kq}{r}$
Electric potential of a system of point charges	$V_P = k\sum_{1}^{N}\frac{q_i}{r_i}$
Electric dipole moment	$\vec{\mathbf{p}} = q\vec{\mathbf{d}}$
Electric potential due to a dipole	$V_P = k\frac{\vec{\mathbf{p}}\cdot\hat{\mathbf{r}}}{r^2}$
Electric potential of a	

continuous charge distribution	$V_P = k\int \frac{dq}{r}$
Electric field components	$E_x = -\frac{\partial V}{\partial x},\ E_y = -\frac{\partial V}{\partial y},\ E_z = -\frac{\partial V}{\partial z}$
Del operator in Cartesian coordinates	$\vec{\nabla} = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \widehat{\mathbf{k}}\frac{\partial}{\partial z}$
Electric field as gradient of potential	$\overrightarrow{\mathbf{E}} = -\vec{\nabla}V$
Del operator in cylindrical coordinates	$\vec{\nabla} = \hat{\mathbf{r}}\frac{\partial}{\partial r} + \widehat{\varphi}\frac{1}{r}\ \frac{\partial}{\partial \varphi} + \widehat{\mathbf{z}}\frac{\partial}{\partial z}$
Del operator in spherical coordinates	$\vec{\nabla} = \hat{\mathbf{r}}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\ \frac{\partial}{\partial \theta} + \widehat{\varphi}\frac{1}{r\sin\theta}\ \frac{\partial}{\partial \varphi}$

Conceptual Questions

Exercise:

Problem:

Why are the metal support rods for satellite network dishes generally grounded?

Solution:

So that lightning striking them goes into the ground instead of the television equipment.

Exercise:

Problem:

(a) Why are fish reasonably safe in an electrical storm? (b) Why are swimmers nonetheless ordered to get out of the water in the same circumstance?

Exercise:

Problem:

What are the similarities and differences between the processes in a photocopier and an electrostatic precipitator?

Solution:

They both make use of static electricity to stick small particles to another surface. However, the precipitator has to charge a wide variety of particles, and is not designed to make sure they land in a particular place.

Exercise:

Problem:

About what magnitude of potential is used to charge the drum of a photocopy machine? A web search for “xerography” may be of use.

Problems

Exercise:

Problem:

(a) What is the electric field 5.00 m from the center of the terminal of a Van de Graaff with a 3.00-mC charge, noting that the field is equivalent to that of a point charge at the center of the terminal? (b) At this distance, what force does the field exert on a $2.00\text{-}\mu\text{C}$ charge on the Van de Graaff’s belt?

Exercise:

Problem:

(a) What is the direction and magnitude of an electric field that supports the weight of a free electron near the surface of Earth? (b) Discuss what the small value for this field implies regarding the relative strength of the gravitational and electrostatic forces.

Solution:

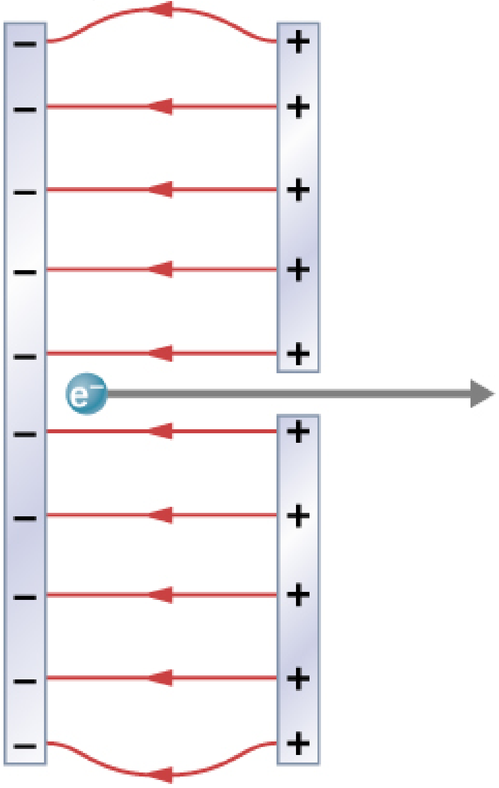
a. $F = 5.58 \times 10^{-11}\ \mathrm{N/C}$;

The electric field is towards the surface of Earth. b. The coulomb force is much stronger than gravity.

Exercise:

Problem:

A simple and common technique for accelerating electrons is shown in [link], where there is a uniform electric field between two plates. Electrons are released, usually from a hot filament, near the negative plate, and there is a small hole in the positive plate that allows the electrons to continue moving. (a) Calculate the acceleration of the electron if the field strength is $2.50 \times 10^{4}\ \mathrm{N/C}$. (b) Explain why the electron will not be pulled back to the positive plate once it moves through the hole.

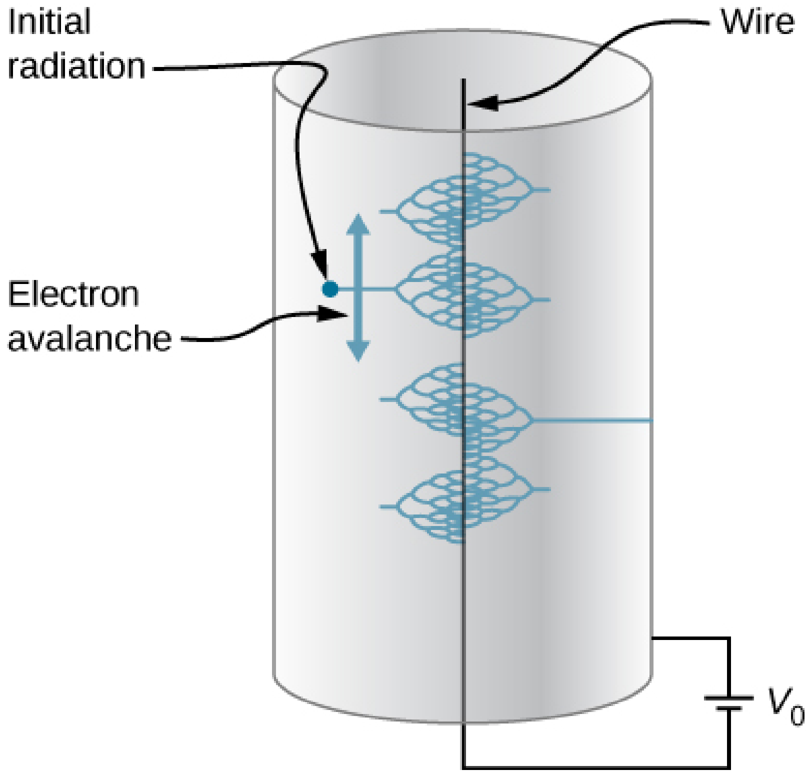


Parallel conducting plates with opposite charges on them create a relatively uniform electric field used to accelerate electrons to to accelerate electrons to the right. Those that go through the hole can be used to make a TV or computer screen glow or to produce X- rays.

Exercise:

Problem:

In a Geiger counter, a thin metallic wire at the center of a metallic tube is kept at a high voltage with respect to the metal tube. Ionizing radiation entering the tube knocks electrons off gas molecules or sides of the tube that then accelerate towards the center wire, knocking off even more electrons. This process eventually leads to an avalanche that is detectable as a current. A particular Geiger counter has a tube of radius R and the inner wire of radius a is at a potential of V_0 volts with respect to the outer metal tube. Consider a point P at a distance s from the center wire and far away from the ends. (a) Find a formula for the electric field at a point P inside using the infinite wire approximation. (b) Find a formula for the electric potential at a point P inside. (c) Use $V_0 = 900\,\text{V}, a = 3.00\,\text{mm}, R = 2.00\,\text{cm}$, and find the value of the electric field at a point 1.00 cm from the center.



Solution:

We know from the Gauss's law chapter that the electric field for an infinite line charge is $\vec{\mathbf{E}}_P = 2k\lambda\frac{1}{s}\hat{\mathbf{s}}$, and from earlier in this chapter that the potential of a wire-cylinder system of this sort is $V_P = -2k\lambda\ln\frac{s_P}{R}$ by integration. We are not given λ, but we are given a fixed V_0; thus, we know that $V_0 = -2k\lambda\ln\frac{a}{R}$ and hence $\lambda = -\frac{V_0}{2k\ln\left(\frac{a}{R}\right)}$. We may substitute this back in to find a. $\vec{\mathbf{E}}_P = -\frac{V_0}{\ln\left(\frac{a}{R}\right)}\ \frac{1}{s}\hat{\mathbf{s}}$; b. $V_P = V_0\frac{\ln\left(\frac{s_P}{R}\right)}{\ln\left(\frac{a}{R}\right)}$; c. $4.74\ \times\ 10^4\,\mathrm{N/C}$

Exercise:

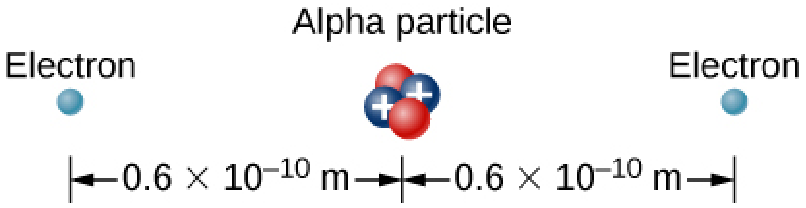
Problem:

The practical limit to an electric field in air is about $3.00\ \times\ 10^{6}\ \text{N/C}$. Above this strength, sparking takes place because air begins to ionize. (a) At this electric field strength, how far would a proton travel before hitting the speed of light (ignore relativistic effects)? (b) Is it practical to leave air in particle accelerators?

Exercise:

Problem:

To form a helium atom, an alpha particle that contains two protons and two neutrons is fixed at one location, and two electrons are brought in from far away, one at a time. The first electron is placed at $0.600\ \times\ 10^{-10}\ \text{m}$ from the alpha particle and held there while the second electron is brought to $0.600\ \times\ 10^{-10}\ \text{m}$ from the alpha particle on the other side from the first electron. See the final configuration below. (a) How much work is done in each step? (b) What is the electrostatic energy of the alpha particle and two electrons in the final configuration?



Solution:

a. $\begin{array}{rcl} U_1 & = & 7.68\ \times\ 10^{-18}\ \text{J} \\ U_2 & = & 5.76\ \times\ 10^{-18}\ \text{J} \end{array}$;

b. $U_1 + U_2 = -1.34\ \times\ 10^{-17}\ \text{J}$

Exercise:

Problem:

Find the electrostatic energy of eight equal charges $(+3\ \mu\text{C})$ each fixed at the corners of a cube of side 2 cm.

Exercise:

Problem:

The probability of fusion occurring is greatly enhanced when appropriate nuclei are brought close together, but mutual Coulomb repulsion must be overcome. This can be done using the kinetic energy of high-temperature gas ions or by accelerating the nuclei toward one another. (a) Calculate the potential energy of two singly charged nuclei separated by $1.00\ \times\ 10^{-12}\,\text{m}$. (b) At what temperature will atoms of a gas have an average kinetic energy equal to this needed electrical potential energy?

Solution:

a. $U = 2.30\ \times\ 10^{-16}\,\text{J}$;

b. $\bar{K} = \frac{3}{2}kT \rightarrow T = 1.11\ \times\ 10^{7}\,\text{K}$

Exercise:

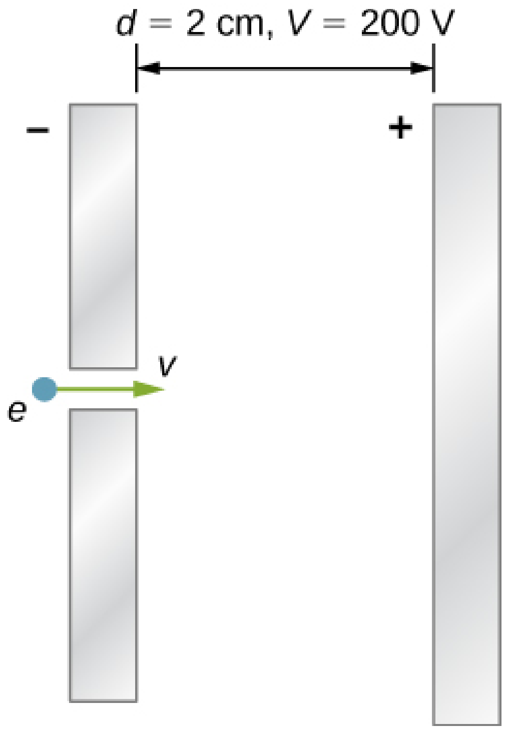
Problem:

A bare helium nucleus has two positive charges and a mass of $6.64\ \times\ 10^{-27}\,\text{kg}$. (a) Calculate its kinetic energy in joules at 2.00% of the speed of light. (b) What is this in electron-volts? (c) What voltage would be needed to obtain this energy?

Exercise:

Problem:

An electron enters a region between two large parallel plates made of aluminum separated by a distance of 2.0 cm and kept at a potential difference of 200 V. The electron enters through a small hole in the negative plate and moves toward the positive plate. At the time the electron is near the negative plate, its speed is $4.0\ \times\ 10^{5}\,\text{m/s}$. Assume the electric field between the plates to be uniform, and find the speed of electron at (a) 0.10 cm, (b) 0.50 cm, (c) 1.0 cm, and (d) 1.5 cm from the negative plate, and (e) immediately before it hits the positive plate.



Solution:

a. 1.9×10^6 m/s; b. 4.2×10^6 m/s; c. 5.9×10^6 m/s; d. 7.3×10^6 m/s; e. 8.4×10^6 m/s

Exercise:

Problem:

How far apart are two conducting plates that have an electric field strength of 4.50×10^3 V/m between them, if their potential difference is 15.0 kV?

Exercise:

Problem:

(a) Will the electric field strength between two parallel conducting plates exceed the breakdown strength of dry air, which is 3.00×10^6 V/m, if the plates are separated by 2.00 mm and a potential difference of 5.0×10^3 V is applied? (b) How close together can the plates be with this applied voltage?

Solution:

a.

$E = 2.5 \times 10^6\ \text{V/m} < 3 \times 10^6\ \text{V/m}$
No, the field strength is smaller than the breakdown strength for air.
;
b. $d = 1.7\,\text{mm}$

Exercise:

Problem:

Membrane walls of living cells have surprisingly large electric fields across them due to separation of ions. What is the voltage across an 8.00-nm-thick membrane if the electric field strength across it is 5.50 MV/m? You may assume a uniform electric field.

Exercise:

Problem:

A double charged ion is accelerated to an energy of 32.0 keV by the electric field between two parallel conducting plates separated by 2.00 cm. What is the electric field strength between the plates?

Solution:

$$\begin{aligned} K_\text{f} &= qV_\text{AB} = qEd \rightarrow \\ E &= 8.00 \times 10^5\ \text{V/m} \end{aligned}$$

Exercise:

Problem:

The temperature near the center of the Sun is thought to be 15 million degrees Celsius $(1.5 \times 10^7\ {}^\circ\text{C})$ (or kelvin). Through what voltage must a singly charged ion be accelerated to have the same energy as the average kinetic energy of ions at this temperature?

Exercise:

Problem:

A lightning bolt strikes a tree, moving 20.0 C of charge through a potential difference of $1.00 \times 10^2 \text{ MV}$. (a) What energy was dissipated? (b) What mass of water could be raised from 15 °C to the boiling point and then boiled by this energy? (c) Discuss the damage that could be caused to the tree by the expansion of the boiling steam.

Solution:

a. $\text{Energy} = 2.00 \times 10^9 \text{ J}$;

b. $\begin{array}{rcl} Q & = & m\left(c\Delta T + L_\nabla\right) \\ m & = & 766\,\text{kg} \end{array}$;

c. The expansion of the steam upon boiling can literally blow the tree apart.

Exercise:

Problem:

What is the potential 0.530×10^{-10} m from a proton (the average distance between the proton and electron in a hydrogen atom)?

Exercise:

Problem:

(a) A sphere has a surface uniformly charged with 1.00 C. At what distance from its center is the potential 5.00 MV? (b) What does your answer imply about the practical aspect of isolating such a large charge?

Solution:

a. $V = \frac{kQ}{r} \rightarrow r = 1.80\,\text{km}$; b. A 1-C charge is a very large amount of charge; a sphere of 1.80 km is impractical.

Exercise:

Problem:

What are the sign and magnitude of a point charge that produces a potential of –2.00 V at a distance of 1.00 mm?

Exercise:

Problem:

In one of the classic nuclear physics experiments at the beginning of the twentieth century, an alpha particle was accelerated toward a gold nucleus, and its path was substantially deflected by the Coulomb interaction. If the energy of the doubly charged alpha nucleus was 5.00 MeV, how close to the gold nucleus (79 protons) could it come before being deflected?

Solution:

The alpha particle approaches the gold nucleus until its original energy is converted to potential energy. $5.00\,\text{MeV} = 8.00\,\times\,10^{-13}\,\text{J}$, so

$$\begin{aligned} E_0 &= \frac{qkQ}{r} \rightarrow \\ r &= 4.54\,\times\,10^{-14}\,\text{m} \end{aligned}$$

(Size of gold nucleus is about $7\,\times\,10^{-15}\,\text{m}$).

Additional Problems

Exercise:

Problem:

A 12.0-V battery-operated bottle warmer heats 50.0 g of glass, $2.50\,\times\,10^{2}$ g of baby formula, and $2.00\,\times\,10^{2}$ g of aluminum from $20.0\,^{\circ}\text{C}$ to $90.0\,^{\circ}\text{C}$. (a) How much charge is moved by the battery? (b) How many electrons per second flow if it takes 5.00 min to warm the formula? (*Hint:* Assume that the specific heat of baby formula is about the same as the specific heat of water.)

Exercise:

Problem:

A battery-operated car uses a 12.0-V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a 2.00×10^2-m high hill, and finally cause it to travel at a constant 25.0 m/s while climbing with 5.00×10^2-N force for an hour.

Solution:

$$\begin{array}{rcl} E_{\text{tot}} & = & 4.67 \times 10^7 \text{ J} \\ E_{\text{tot}} & = & qV \rightarrow q = \frac{E_{\text{tot}}}{V} = 3.89 \times 10^6 \text{ C} \end{array}$$

Exercise:

Problem:

(a) Find the voltage near a 10.0 cm diameter metal sphere that has 8.00 C of excess positive charge on it. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem:

A uniformly charged half-ring of radius 10 cm is placed on a nonconducting table. It is found that 3.0 cm above the center of the half-ring the potential is –3.0 V with respect to zero potential at infinity. How much charge is in the half-ring?

Solution:

$$V_P = k\frac{q_{\text{tot}}}{\sqrt{z^2+R^2}} \rightarrow q_{\text{tot}} = -3.5 \times 10^{-11} \text{ C}$$

Exercise:

Problem:

A glass ring of radius 5.0 cm is painted with a charged paint such that the charge density around the ring varies continuously given by the following function of the polar angle $\theta, \lambda = \left(3.0 \times 10^{-6}\ \mathrm{C/m}\right) \cos^2\theta$. Find the potential at a point 15 cm above the center.

Exercise:

Problem:

A CD disk of radius ($R = 3.0\ \mathrm{cm}$) is sprayed with a charged paint so that the charge varies continually with radial distance *r* from the center in the following manner: $\sigma = -\,(6.0\ \mathrm{C/m})r/R$.

Find the potential at a point 4 cm above the center.

Solution:

$V_P = -2.2\ \mathrm{GV}$

Exercise:

Problem:

(a) What is the final speed of an electron accelerated from rest through a voltage of 25.0 MV by a negatively charged Van de Graff terminal? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem:

A large metal plate is charged uniformly to a density of $\sigma = 2.0 \times 10^{-9}\ \mathrm{C/m^2}$. How far apart are the equipotential surfaces that represent a potential difference of 25 V?

Solution:

Recall from the previous chapter that the electric field $E_P = \frac{\sigma_0}{2\varepsilon_0}$ is uniform throughout space, and that for uniform fields we have

$E = -\frac{\Delta V}{\Delta z}$ for the relation. Thus, we get $\frac{\sigma}{2\varepsilon_0} = \frac{\Delta V}{\Delta z} \rightarrow \Delta z = 0.22\,\text{m}$ for the distance between 25-V equipotentials.

Exercise:

Problem:

Your friend gets really excited by the idea of making a lightning rod or maybe just a sparking toy by connecting two spheres as shown in [link], and making R_2 so small that the electric field is greater than the dielectric strength of air, just from the usual 150 V/m electric field near the surface of the Earth. If R_1 is 10 cm, how small does R_2 need to be, and does this seem practical? (*Hint:* recall the calculation for electric field at the surface of a conductor from Gauss's Law.)

Exercise:

Problem:

(a) Find $x \;>>\; L$ limit of the potential of a finite uniformly charged rod and show that it coincides with that of a point charge formula. (b) Why would you expect this result?

Solution:

a. Take the result from [link], divide both the numerator and the denominator by *x*, take the limit of that, and then apply a Taylor expansion to the resulting log to get: $V_P \approx k\lambda\frac{L}{x}$; b. which is the result we expect, because at great distances, this should look like a point charge of $q = \lambda L$

Exercise:

Problem:

A small spherical pith ball of radius 0.50 cm is painted with a silver paint and then $-10\;\mu\text{C}$ of charge is placed on it. The charged pith ball is put at the center of a gold spherical shell of inner radius 2.0 cm and outer radius 2.2 cm. (a) Find the electric potential of the gold shell with respect to zero potential at infinity. (b) How much charge should you put on the gold shell if you want to make its potential 100 V?

Exercise:

Problem:

Two parallel conducting plates, each of cross-sectional area $400\ \text{cm}^2$, are 2.0 cm apart and uncharged. If $1.0\ \times\ 10^{12}$ electrons are transferred from one plate to the other, (a) what is the potential difference between the plates? (b) What is the potential difference between the positive plate and a point 1.25 cm from it that is between the plates?

Solution:

a. $V = 9.0\ \times\ 10^3\ \text{V}$; b. $-9.0\ \times\ 10^3\ \text{V}\left(\frac{1.25\ \text{cm}}{2.0\ \text{cm}}\right) = -5.7\ \times\ 10^3\ \text{V}$

Exercise:

Problem:

A point charge of $q = 5.0\ \times\ 10^{-8}\ \text{C}$ is placed at the center of an uncharged spherical conducting shell of inner radius 6.0 cm and outer radius 9.0 cm. Find the electric potential at (a) $r = 4.0\ \text{cm}$, (b) $r = 8.0\ \text{cm}$, (c) $r = 12.0\ \text{cm}$.

Exercise:

Problem:

Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface. (a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center? (b) What acceleration will the field produce on a free electron near Earth's surface? (c) What mass object with a single extra electron will have its weight supported by this field?

Solution:

a. $E = \frac{KQ}{r^2} \rightarrow Q = -6.76\ \times\ 10^5\ \text{C}$;

b. $\begin{array}{rcl} F & = & ma = qE \rightarrow \\ a & = & \frac{qE}{m} = 2.63\ \times\ 10^{13}\ \text{m/s}^2\ (\text{upwards}) \end{array}$;

c. $F = -mg = qE \rightarrow m = \frac{-qE}{g} = 2.45 \times 10^{-18}$ kg

Exercise:

Problem:

Point charges of $25.0\ \mu\text{C}$ and $45.0\ \mu\text{C}$ are placed 0.500 m apart.

(a) At what point along the line between them is the electric field zero?

(b) What is the electric field halfway between them?

Exercise:

Problem:

What can you say about two charges q_1 and q_2, if the electric field one-fourth of the way from q_1 to q_2 is zero?

Solution:

If the electric field is zero ¼ from the way of q_1 and q_2, then we know from $E = k\frac{Q}{r^2}$ that $|E_1| = |E_2| \rightarrow \frac{Kq_1}{x^2} = \frac{Kq_2}{(3x)^2}$ so that $\frac{q_2}{q_1} = \frac{(3x)^2}{x^2} = 9;$ the charge q_2 is 9 times larger than q_1.

Exercise:

Problem:

Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is 0.530×10^{-10} m. You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.

Exercise:

Problem:

An electron has an initial velocity of 5.00×10^6 m/s in a uniform 2.00×10^5-N/C electric field. The field accelerates the electron in the direction opposite to its initial velocity. (a) What is the direction of the electric field? (b) How far does the electron travel before coming to rest? (c) How long does it take the electron to come to rest? (d) What is the electron’s velocity when it returns to its starting point?

Solution:

a. The field is in the direction of the electron’s initial velocity.

b.

$$\begin{aligned} v^2 &= v_0^2 + 2ax \rightarrow x = -\frac{v_0^2}{2a}\,(v = 0).\ \text{Also, } F = ma = qE \rightarrow a = \frac{qE}{m}, \\ x &= 3.56 \times 10^{-4}\ \text{m;} \end{aligned}$$

c.

$$\begin{aligned} v_2 &= v_0 + at \rightarrow t = -\frac{v_0 m}{qE}\,(v = 0), \\ \therefore t &= 1.42 \times 10^{-10}\ \text{s;} \end{aligned}$$

d. $v = -\left(\frac{2qEx}{m}\right)^{1/2} - 5.00 \times 10^6\ \text{m/s}$ (opposite its initial velocity)

Challenge Problems

Exercise:

Problem:

Three Na^+ and three Cl^- ions are placed alternately and equally spaced around a circle of radius 50 nm. Find the electrostatic energy stored.

Exercise:

Problem:

Look up (presumably online, or by dismantling an old device and making measurements) the magnitude of the potential deflection plates (and the space between them) in an ink jet printer. Then look up the speed with which the ink comes out the nozzle. Can you calculate the typical mass of an ink drop?

Solution:

Answers will vary. This appears to be proprietary information, and ridiculously difficult to find. Speeds will be 20 m/s or less, and there are claims of ~10^{-7} grams for the mass of a drop.

Exercise:

Problem:

Use the electric field of a finite sphere with constant volume charge density to calculate the electric potential, throughout space. Then check your results by calculating the electric field from the potential.

Exercise:

Problem:

Calculate the electric field of a dipole throughout space from the potential.

Solution:

Apply $\overrightarrow{\mathbf{E}} = -\vec{\nabla}V$ with $\vec{\nabla} = \hat{\mathbf{r}}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\ \frac{\partial}{\partial \theta} + \widehat{\varphi}\frac{1}{r\sin\theta}\ \frac{\partial}{\partial \varphi}$ to the potential calculated earlier, $V_P = k\frac{\vec{\mathbf{p}}\cdot\hat{\mathbf{r}}}{r^2}$ with $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$, and assume that the axis of the dipole is aligned with the *z*-axis of the coordinate system. Thus, the potential is $V_P = k\frac{q\vec{\mathbf{d}}\cdot\hat{\mathbf{r}}}{r^2} = k\frac{qd\cos\theta}{r^2}$.

$\overrightarrow{\mathbf{E}} = 2kqd\left(\frac{\cos\theta}{r^3}\right)\hat{\mathbf{r}} + kqd\left(\frac{\sin\theta}{r^3}\right)\hat{\theta}$

Glossary

electrostatic precipitators
filters that apply charges to particles in the air, then attract those charges to a filter, removing them from the airstream

ink jet printer

small ink droplets sprayed with an electric charge are controlled by electrostatic plates to create images on paper

photoconductor
substance that is an insulator until it is exposed to light, when it becomes a conductor

Van de Graaff generator
machine that produces a large amount of excess charge, used for experiments with high voltage

xerography
dry copying process based on electrostatics

Introduction class="introduction"

The tree-like branch patterns in this clear Plexiglas® block are known as a Lichtenberg figure, named for the German physicist Georg Christof Lichtenberg (1742–1799), who was the first to study these patterns. The “branches” are created by the dielectric breakdown produced by a strong electric field. (credit: modification of work

by Bert
Hickman)

Capacitors are important components of electrical circuits in many electronic devices, including pacemakers, cell phones, and computers. In this chapter, we study their properties, and, over the next few chapters, we examine their function in combination with other circuit elements. By themselves, capacitors are often used to store electrical energy and release it when needed; with other circuit components, capacitors often act as part of a filter that allows some electrical signals to pass while blocking others. You can see why capacitors are considered one of the fundamental components of electrical circuits.

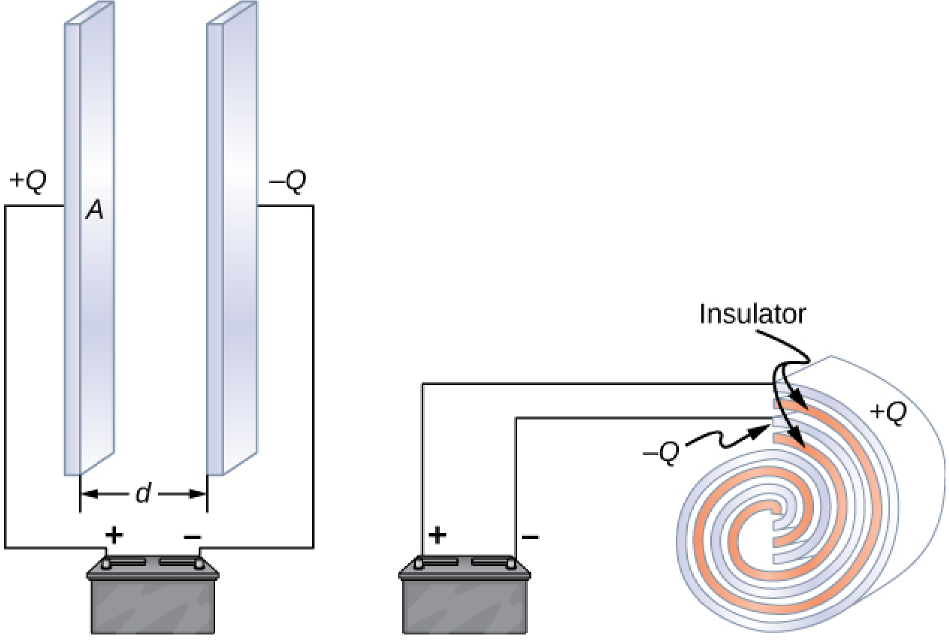
Capacitors and Capacitance

By the end of this section, you will be able to:

- Explain the concepts of a capacitor and its capacitance
- Describe how to evaluate the capacitance of a system of conductors

A **capacitor** is a device used to store electrical charge and electrical energy. It consists of at least two electrical conductors separated by a distance. (Note that such electrical conductors are sometimes referred to as “electrodes,” but more correctly, they are “capacitor plates.”) The space between capacitors may simply be a vacuum, and, in that case, a capacitor is then known as a “vacuum capacitor.” However, the space is usually filled with an insulating material known as a **dielectric**. (You will learn more about dielectrics in the sections on dielectrics later in this chapter.) The amount of storage in a capacitor is determined by a property called *capacitance*, which you will learn more about a bit later in this section.

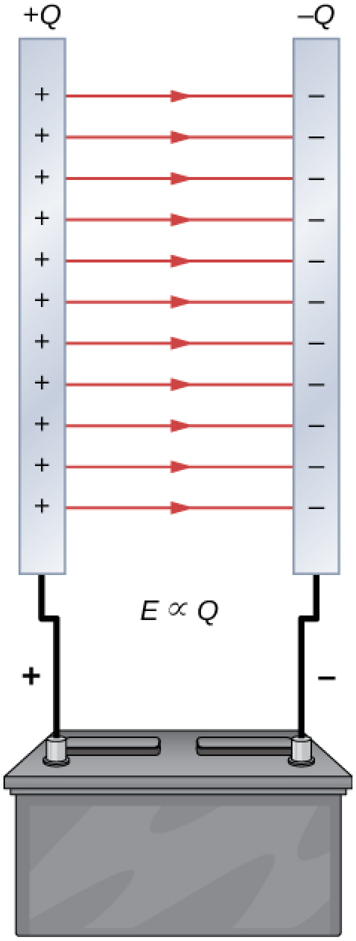
Capacitors have applications ranging from filtering static from radio reception to energy storage in heart defibrillators. Typically, commercial capacitors have two conducting parts close to one another but not touching, such as those in [link]. Most of the time, a dielectric is used between the two plates. When battery terminals are connected to an initially uncharged capacitor, the battery potential moves a small amount of charge of magnitude Q from the positive plate to the negative plate. The capacitor remains neutral overall, but with charges $+Q$ and $-Q$ residing on opposite plates.



(a) (b)

Both capacitors shown here were initially uncharged before being connected to a battery. They now have charges of $+Q$ and $-Q$ (respectively) on their plates. (a) A parallel-plate capacitor consists of two plates of opposite charge with area *A* separated by distance *d*. (b) A rolled capacitor has a dielectric material between its two conducting sheets (plates).

A system composed of two identical parallel-conducting plates separated by a distance is called a **parallel-plate capacitor** ([link]). The magnitude of the electrical field in the space between the parallel plates is $E=\sigma/\varepsilon_0$, where σ denotes the surface charge density on one plate (recall that σ is the charge Q per the surface area A). Thus, the magnitude of the field is directly proportional to Q.



The charge separation in a capacitor shows that the charges remain on the surfaces of the capacitor plates. Electrical field lines in a parallel-plate capacitor begin with positive charges and end with negative charges. The magnitude of the electrical field in the space between the plates is in direct proportion to the amount of charge on the capacitor.

Capacitors with different physical characteristics (such as shape and size of their plates) store different amounts of charge for the same applied voltage V across their plates. The **capacitance** C of a capacitor is defined as the ratio of the maximum charge Q that can be stored in a capacitor to the applied voltage V across its plates. In other words, capacitance is the largest amount of charge per volt that can be stored on the device:

Note:
Equation:

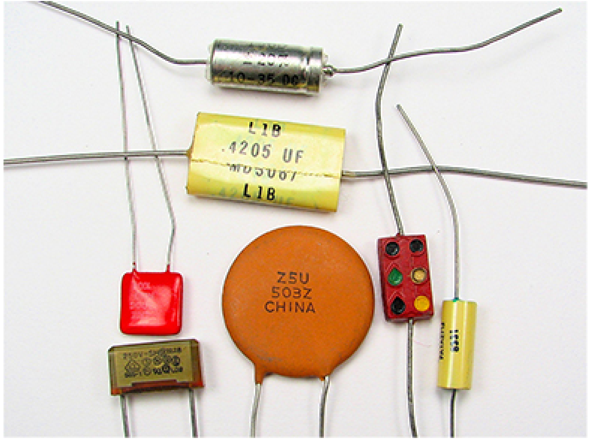
$$C = \frac{Q}{V}.$$

The SI unit of capacitance is the farad (F), named after Michael Faraday (1791–1867). Since capacitance is the charge per unit voltage, one farad is one coulomb per one volt, or

Equation:

$$1\,\text{F} = \frac{1\,\text{C}}{1\,\text{V}}.$$

By definition, a 1.0-F capacitor is able to store 1.0 C of charge (a very large amount of charge) when the potential difference between its plates is only 1.0 V. One farad is therefore a very large capacitance. Typical capacitance values range from picofarads $(1\,\text{pF} = 10^{-12}\,\text{F})$ to millifarads $(1\,\text{mF} = 10^{-3}\,\text{F})$, which also includes microfarads $(1\,\mu\text{F} = 10^{-6}\,\text{F})$. Capacitors can be produced in various shapes and sizes ([link]).



These are some typical capacitors used in electronic devices. A capacitor’s size is not necessarily related to its capacitance value. (credit: Windell Oskay)

Calculation of Capacitance

We can calculate the capacitance of a pair of conductors with the standard approach that follows.

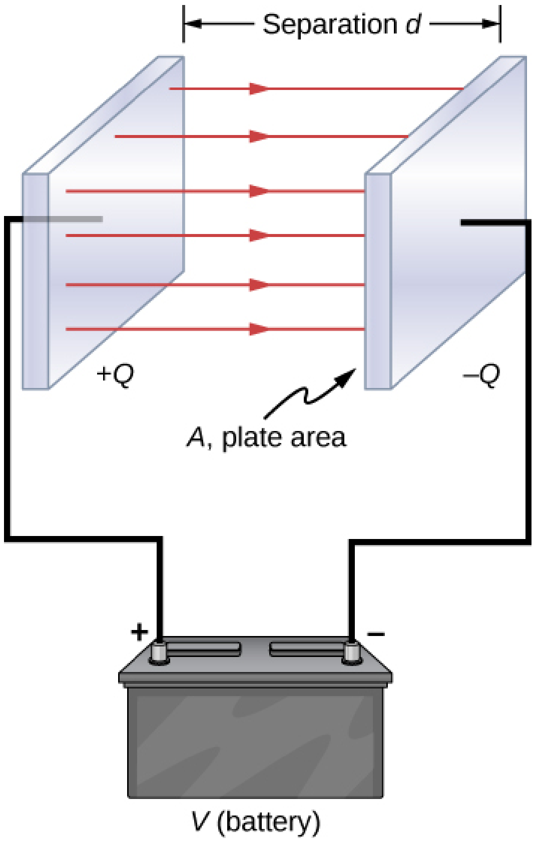
Note:
Problem-Solving Strategy: Calculating Capacitance

1. Assume that the capacitor has a charge Q.
2. Determine the electrical field $\vec{\mathbf{E}}$ between the conductors. If symmetry is present in the arrangement of conductors, you may be able to use Gauss's law for this calculation.
3. Find the potential difference between the conductors from
   Equation:
   $$V_B - V_A = -\int_A^B \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}},$$
   where the path of integration leads from one conductor to the other. The magnitude of the potential difference is then $V = |V_B - V_A|$.
4. With V known, obtain the capacitance directly from [link].

To show how this procedure works, we now calculate the capacitances of parallel-plate, spherical, and cylindrical capacitors. In all cases, we assume vacuum capacitors (empty capacitors) with no dielectric substance in the space between conductors.

Parallel-Plate Capacitor

The parallel-plate capacitor ([link]) has two identical conducting plates, each having a surface area A, separated by a distance d. When a voltage V is applied to the capacitor, it stores a charge Q, as shown. We can see how its capacitance may depend on A and d by considering characteristics of the Coulomb force. We know that force between the charges increases with charge values and decreases with the distance between them. We should expect that the bigger the plates are, the more charge they can store. Thus, C should be greater for a larger value of A. Similarly, the closer the plates are together, the greater the attraction of the opposite charges on them. Therefore, C should be greater for a smaller d.



In a parallel-plate capacitor with plates separated by a distance *d*, each plate has the same surface area *A*.

We define the surface charge density σ on the plates as

Equation:

$$\sigma = \frac{Q}{A}.$$

We know from previous chapters that when *d* is small, the electrical field between the plates is fairly uniform (ignoring edge effects) and that its magnitude is given by

Equation:

$$E = \frac{\sigma}{\varepsilon_0},$$

where the constant ε_0 is the permittivity of free space, $\varepsilon_0 = 8.85\ \times\ 10^{-12}\ \mathrm{F/m}$. The SI unit of F/m is equivalent to $\mathrm{C^2/N \cdot m^2}$. Since the electrical field $\vec{\mathbf{E}}$ between the plates is uniform, the potential difference between the plates is

Equation:

$$V = Ed = \frac{\sigma d}{\varepsilon_0} = \frac{Qd}{\varepsilon_0 A}.$$

Therefore [link] gives the capacitance of a parallel-plate capacitor as

Note:
Equation:

$$C = \frac{Q}{V} = \frac{Q}{Qd/\varepsilon_0 A} = \varepsilon_0 \frac{A}{d}.$$

Notice from this equation that capacitance is a function *only of the geometry* and what material fills the space between the plates (in this case, vacuum) of this capacitor. In fact, this is true not only for a parallel-plate capacitor, but for all capacitors: The capacitance is independent of Q or V. If the charge changes, the potential changes correspondingly so that Q/V remains constant.

Example:
Capacitance and Charge Stored in a Parallel-Plate Capacitor
(a) What is the capacitance of an empty parallel-plate capacitor with metal plates that each have an area of $1.00\ \text{m}^2$, separated by 1.00 mm? (b) How much charge is stored in this capacitor if a voltage of 3.00×10^3 V is applied to it?
Strategy
Finding the capacitance C is a straightforward application of [link]. Once we find C, we can find the charge stored by using [link].
Solution

a. Entering the given values into [link] yields
   Equation:

$$C = \varepsilon_0 \frac{A}{d} = \left(8.85 \times 10^{-12} \frac{\text{F}}{\text{m}}\right) \frac{1.00\,\text{m}^2}{1.00 \times 10^{-3}\,\text{m}} = 8.85 \times 10^{-9}\,\text{F} = 8.85\,\text{nF}.$$

   This small capacitance value indicates how difficult it is to make a device with a large capacitance.
b. Inverting [link] and entering the known values into this equation gives
   Equation:

$$Q = CV = (8.85 \times 10^{-9}\,\text{F})(3.00 \times 10^3\,\text{V}) = 26.6\,\mu\text{C}.$$

Significance

This charge is only slightly greater than those found in typical static electricity applications. Since air breaks down (becomes conductive) at an electrical field strength of about 3.0 MV/m, no more charge can be stored on this capacitor by increasing the voltage.

Example:
A 1-F Parallel-Plate Capacitor
Suppose you wish to construct a parallel-plate capacitor with a capacitance of 1.0 F. What area must you use for each plate if the plates are separated by 1.0 mm?
Solution
Rearranging [link], we obtain
Equation:

$$A = \frac{Cd}{\varepsilon_0} = \frac{(1.0\,\text{F})(1.0\,\times\,10^{-3}\,\text{m})}{8.85\,\times\,10^{-12}\,\text{F/m}} = 1.1\,\times\,10^{8}\,\text{m}^2.$$

Each square plate would have to be 10 km across. It used to be a common prank to ask a student to go to the laboratory stockroom and request a 1-F parallel-plate capacitor, until stockroom attendants got tired of the joke.

Note:
Exercise:

Problem:

Check Your Understanding The capacitance of a parallel-plate capacitor is 2.0 pF. If the area of each plate is $2.4\,\text{cm}^2$, what is the plate separation?

Solution:

$1.1\,\times\,10^{-3}\,\text{m}$

Note:
Exercise:

Problem: Check Your Understanding Verify that σ/V and ε_0/d have the same physical units.

Spherical Capacitor

A spherical capacitor is another set of conductors whose capacitance can be easily determined ([link]). It consists of two concentric conducting spherical shells of radii R_1 (inner shell) and R_2 (outer shell). The shells are given equal and opposite charges $+Q$ and $-Q$, respectively. From symmetry, the electrical field between the shells is directed radially outward. We can obtain the magnitude of the field by applying Gauss’s law over a spherical Gaussian surface of radius *r* concentric with the shells. The enclosed charge is $+Q$; therefore we have

Equation:

$$\oint_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}} dA = E(4\pi r^2) = \frac{Q}{\varepsilon_0}.$$

Thus, the electrical field between the conductors is

Equation:

$$\vec{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} \hat{\mathbf{r}}.$$

We substitute this $\vec{\mathbf{E}}$ into [link] and integrate along a radial path between the shells:

Equation:

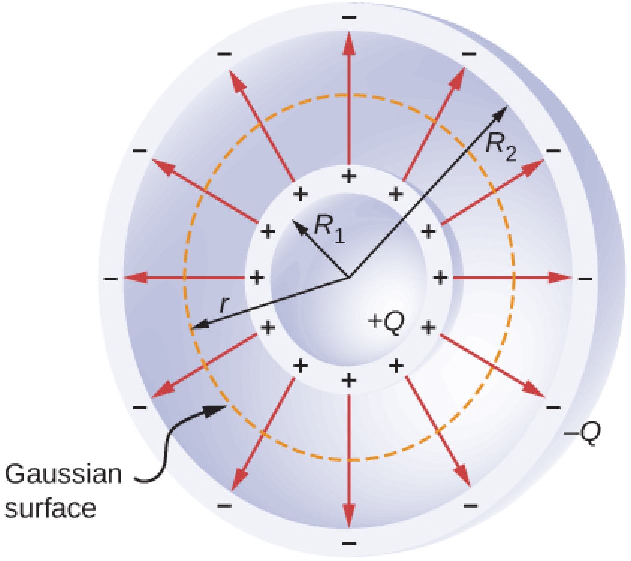
$$V = \int_{R_1}^{R_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = \int_{R_1}^{R_2} \left(\frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} \hat{\mathbf{r}} \right) \cdot (\hat{\mathbf{r}}\, dr) = \frac{Q}{4\pi\varepsilon_0} \int_{R_1}^{R_2} \frac{dr}{r^2} = \frac{Q}{4\pi\varepsilon_0} \left(\frac{1}{R_1} - \frac{1}{R_2} \right).$$

In this equation, the potential difference between the plates is $V = -(V_2 - V_1) = V_1 - V_2$. We substitute this result into [link] to find the capacitance of a spherical capacitor:

Note:

Equation:

$$C = \frac{Q}{V} = 4\pi\varepsilon_0 \frac{R_1 R_2}{R_2 - R_1}.$$



A spherical capacitor consists of two concentric conducting spheres. Note that the charges on a conductor reside on its surface.

Example:

Capacitance of an Isolated Sphere

Calculate the capacitance of a single isolated conducting sphere of radius R_1 and compare it with [link] in the limit as $R_2 \to \infty$.

Strategy

We assume that the charge on the sphere is Q, and so we follow the four steps outlined earlier. We also assume the other conductor to be a concentric hollow sphere of infinite radius.

Solution

On the outside of an isolated conducting sphere, the electrical field is given by [link]. The magnitude of the potential difference between the surface of an isolated sphere and infinity is

Equation:

$$V = \int_{R_1}^{+\infty} \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = \frac{Q}{4\pi\varepsilon_0} \int_{R_1}^{+\infty} \frac{1}{r^2} \hat{\mathbf{r}} \cdot (\hat{\mathbf{r}}\, dr) = \frac{Q}{4\pi\varepsilon_0} \int_{R_1}^{+\infty} \frac{dr}{r^2} = \frac{1}{4\pi\varepsilon_0} \frac{Q}{R_1}.$$

The capacitance of an isolated sphere is therefore

Equation:

$$C = \frac{Q}{V} = Q\frac{4\pi\varepsilon_0 R_1}{Q} = 4\pi\varepsilon_0 R_1.$$

Significance

The same result can be obtained by taking the limit of [link] as $R_2 \to \infty$. A single isolated sphere is therefore equivalent to a spherical capacitor whose outer shell has an infinitely large radius.

Note:
Exercise:

Problem:

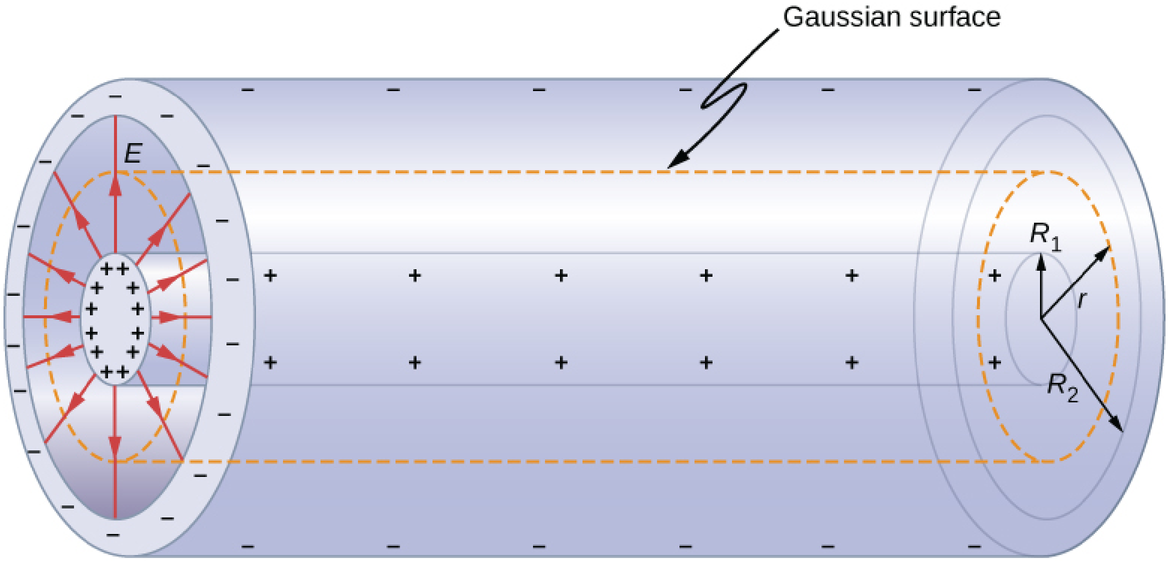
Check Your Understanding The radius of the outer sphere of a spherical capacitor is five times the radius of its inner shell. What are the dimensions of this capacitor if its capacitance is 5.00 pF?

Solution:

3.59 cm, 17.98 cm

Cylindrical Capacitor

A cylindrical capacitor consists of two concentric, conducting cylinders ([link]). The inner cylinder, of radius R_1, may either be a shell or be completely solid. The outer cylinder is a shell of inner radius R_2. We assume that the length of each cylinder is l and that the excess charges $+Q$ and $-Q$ reside on the inner and outer cylinders, respectively.



A cylindrical capacitor consists of two concentric, conducting cylinders. Here, the charge on the outer surface of the inner cylinder is positive (indicated by $+$) and the charge on the inner surface of the outer cylinder is negative (indicated by $-$).

With edge effects ignored, the electrical field between the conductors is directed radially outward from the common axis of the cylinders. Using the Gaussian surface shown in [link], we have

Equation:

$$\oint_S \vec{\mathbf{E}} \cdot \hat{\mathbf{n}} \, dA = E(2\pi r l) = \frac{Q}{\varepsilon_0}.$$

Therefore, the electrical field between the cylinders is
Equation:

$$\vec{\mathbf{E}} = \frac{1}{2\pi\varepsilon_0} \frac{Q}{r\,l} \hat{\mathbf{r}}.$$

Here $\hat{\mathbf{r}}$ is the unit radial vector along the radius of the cylinder. We can substitute into [link] and find the potential difference between the cylinders:
Equation:

$$V = \int_{R_1}^{R_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}}_p = \frac{Q}{2\pi\varepsilon_0\, l} \int_{R_1}^{R_2} \frac{1}{r} \hat{\mathbf{r}} \cdot (\hat{\mathbf{r}}\, dr) = \frac{Q}{2\pi\varepsilon_0\, l} \int_{R_1}^{R_2} \frac{dr}{r} = \frac{Q}{2\pi\varepsilon_0\, l} \ln r|_{R_1}^{R_2} = \frac{Q}{2\pi\varepsilon_0\, l} \ln \frac{R_2}{R_1}.$$

Thus, the capacitance of a cylindrical capacitor is

Note:
Equation:

$$C = \frac{Q}{V} = \frac{2\pi\varepsilon_0\, l}{\ln(R_2/R_1)}.$$

As in other cases, this capacitance depends only on the geometry of the conductor arrangement. An important application of [link] is the determination of the capacitance per unit length of a *coaxial cable*, which is commonly used to transmit time-varying electrical signals. A coaxial cable consists of two concentric, cylindrical conductors separated by an insulating material. (Here, we assume a vacuum between the conductors, but the physics is qualitatively almost the same when the space between the conductors is filled by a dielectric.) This configuration shields the electrical signal propagating down the inner conductor from stray electrical fields external to the cable. Current flows in opposite directions in the inner and the outer conductors, with the outer conductor usually grounded. Now, from [link], the capacitance per unit length of the coaxial cable is given by
Equation:

$$\frac{C}{l} = \frac{2\pi\varepsilon_0}{\ln(R_2/R_1)}.$$

In practical applications, it is important to select specific values of C/l. This can be accomplished with appropriate choices of radii of the conductors and of the insulating material between them.

Note:
Exercise:

Problem:

Check Your Understanding When a cylindrical capacitor is given a charge of 0.500 nC, a potential difference of 20.0 V is measured between the cylinders. (a) What is the capacitance of this system? (b) If the cylinders are 1.0 m long, what is the ratio of their radii?

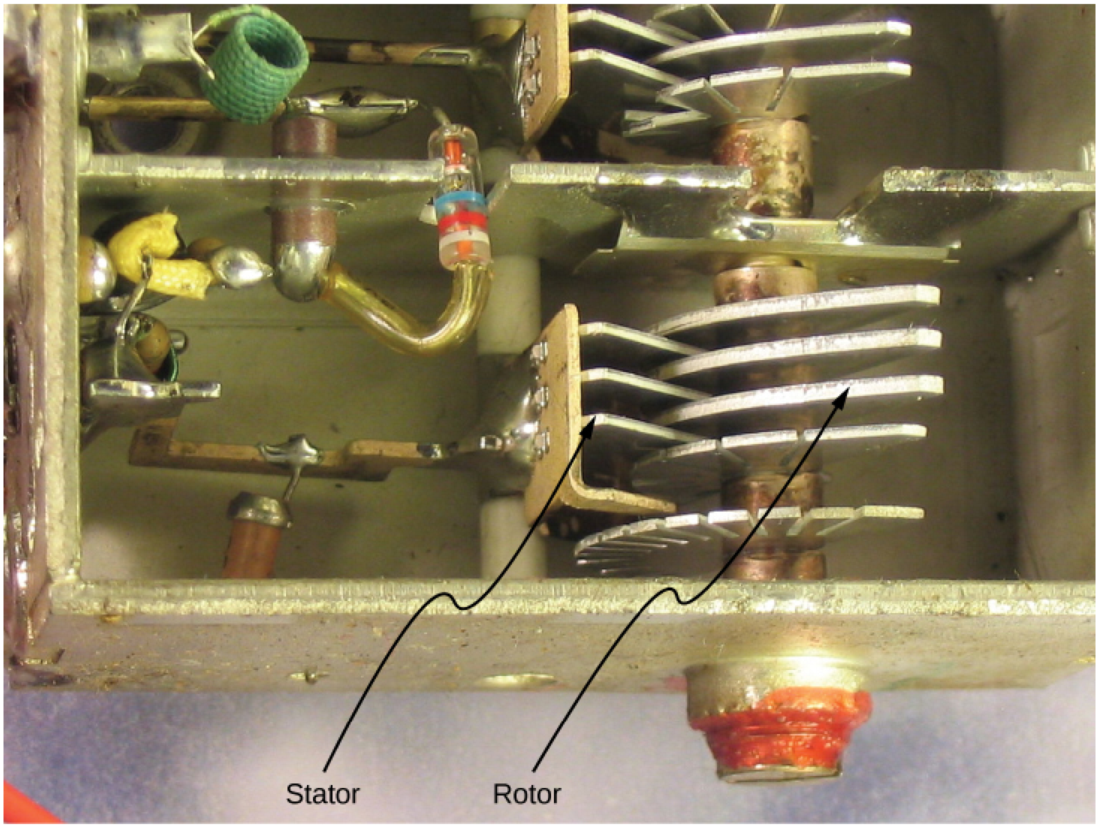
Solution:

a. 25.0 pF; b. 9.2

Several types of practical capacitors are shown in [link]. Common capacitors are often made of two small pieces of metal foil separated by two small pieces of insulation (see [link](b)). The metal foil and insulation are encased in a protective coating, and two metal leads are used for connecting the foils to an external circuit. Some common insulating materials are mica, ceramic, paper, and Teflon™ non-stick coating.

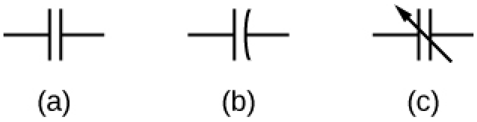
Another popular type of capacitor is an electrolytic capacitor. It consists of an oxidized metal in a conducting paste. The main advantage of an electrolytic capacitor is its high capacitance relative to other common types of capacitors. For example, capacitance of one type of aluminum electrolytic capacitor can be as high as 1.0 F. However, you must be careful when using an electrolytic capacitor in a circuit, because it only functions correctly when the metal foil is at a higher potential than the conducting paste. When reverse polarization occurs, electrolytic action destroys the oxide film. This type of capacitor cannot be connected across an alternating current source, because half of the time, ac voltage would have the wrong polarity, as an alternating current reverses its polarity (see Alternating-Current Circuts on alternating-current circuits).

A variable air capacitor ([link]) has two sets of parallel plates. One set of plates is fixed (indicated as “stator”), and the other set of plates is attached to a shaft that can be rotated (indicated as “rotor”). By turning the shaft, the cross-sectional area in the overlap of the plates can be changed; therefore, the capacitance of this system can be tuned to a desired value. Capacitor tuning has applications in any type of radio transmission and in receiving radio signals from electronic devices. Any time you tune your car radio to your favorite station, think of capacitance.



In a variable air capacitor, capacitance can be tuned by changing the effective area of the plates. (credit: modification of work by Robbie Sproule)

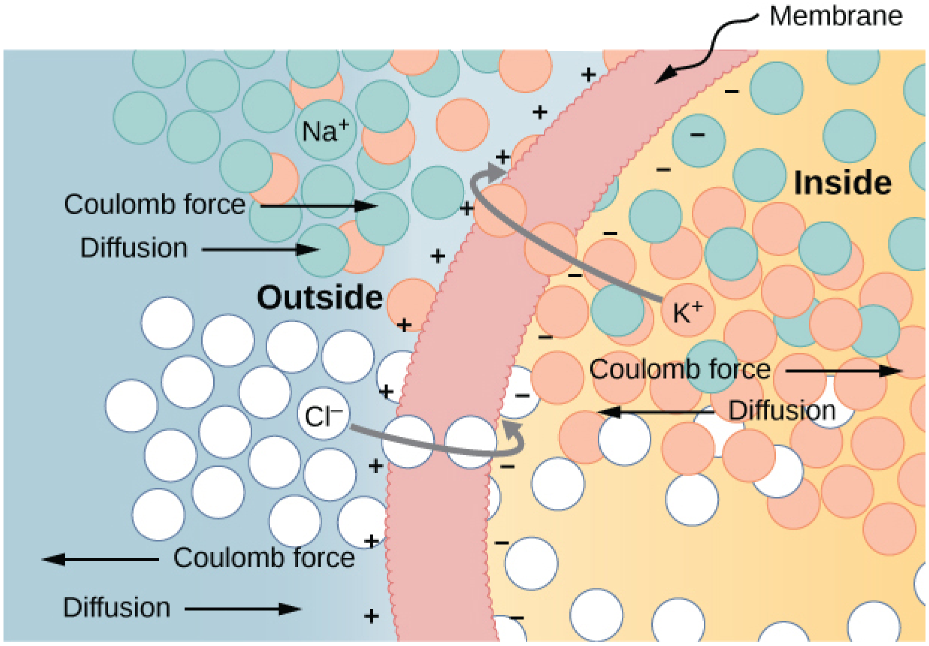
The symbols shown in [link] are circuit representations of various types of capacitors. We generally use the symbol shown in [link](a). The symbol in [link](c) represents a variable-capacitance capacitor. Notice the similarity of these symbols to the symmetry of a parallel-plate capacitor. An electrolytic capacitor is represented by the symbol in part [link](b), where the curved plate indicates the negative terminal.



This shows three different circuit representations of capacitors. The symbol in (a) is the most commonly used one. The symbol in (b) represents an electrolytic capacitor. The symbol in (c) represents a variable-capacitance capacitor.

An interesting applied example of a capacitor model comes from cell biology and deals with the electrical potential in the plasma membrane of a living cell ([link]). Cell membranes separate cells from their surroundings but allow some selected ions to pass in or out of the cell. The potential difference across a membrane is about 70 mV. The cell membrane may be 7 to 10 nm thick. Treating the cell membrane as a nano-sized capacitor, the estimate of the smallest electrical field strength across its 'plates' yields the value $E = \frac{V}{d} = \frac{70 \times 10^{-3}\text{V}}{10 \times 10^{-9}\text{m}} = 7 \times 10^{6}\ \text{V/m} > 3\,\text{MV/m}.$

This magnitude of electrical field is great enough to create an electrical spark in the air.



The semipermeable membrane of a biological cell has different concentrations of ions on its interior surface than on its exterior. Diffusion moves the K^{+} (potassium) and Cl^{-} (chloride) ions in the directions shown, until the Coulomb force halts further transfer. In this way, the exterior of the membrane acquires a positive charge and its interior surface acquires a negative charge, creating a potential difference across the membrane. The membrane is normally impermeable to Na+ (sodium ions).

Note:
Visit the PhET Explorations: Capacitor Lab to explore how a capacitor works. Change the size of the plates and add a dielectric to see the effect on capacitance. Change the voltage and see charges built

up on the plates. Observe the electrical field in the capacitor. Measure the voltage and the electrical field.

Summary

- A capacitor is a device that stores an electrical charge and electrical energy. The amount of charge a vacuum capacitor can store depends on two major factors: the voltage applied and the capacitor’s physical characteristics, such as its size and geometry.
- The capacitance of a capacitor is a parameter that tells us how much charge can be stored in the capacitor per unit potential difference between its plates. Capacitance of a system of conductors depends only on the geometry of their arrangement and physical properties of the insulating material that fills the space between the conductors. The unit of capacitance is the farad, where $1\,\text{F} = 1\,\text{C}/1\,\text{V}$.

Conceptual Questions

Exercise:

Problem:

Does the capacitance of a device depend on the applied voltage? Does the capacitance of a device depend on the charge residing on it?

Solution:

no; yes

Exercise:

Problem:

Would you place the plates of a parallel-plate capacitor closer together or farther apart to increase their capacitance?

Exercise:

Problem: The value of the capacitance is zero if the plates are not charged. True or false?

Solution:

false

Exercise:

Problem:

If the plates of a capacitor have different areas, will they acquire the same charge when the capacitor is connected across a battery?

Exercise:

Problem:

Does the capacitance of a spherical capacitor depend on which sphere is charged positively or negatively?

Solution:

no

Problems

Exercise:

Problem: What charge is stored in a $180.0\text{-}\mu\text{F}$ capacitor when 120.0 V is applied to it?

Solution:

21.6 mC

Exercise:

Problem: Find the charge stored when 5.50 V is applied to an 8.00-pF capacitor.

Exercise:

Problem: Calculate the voltage applied to a $2.00\text{-}\mu\text{F}$ capacitor when it holds $3.10\ \mu\text{C}$ of charge.

Solution:

1.55 V

Exercise:

Problem: What voltage must be applied to an 8.00-nF capacitor to store 0.160 mC of charge?

Exercise:

Problem: What capacitance is needed to store $3.00\ \mu\text{C}$ of charge at a voltage of 120 V?

Solution:

25.0 nF

Exercise:

Problem:

What is the capacitance of a large Van de Graaff generator’s terminal, given that it stores 8.00 mC of charge at a voltage of 12.0 MV?

Exercise:

Problem:

The plates of an empty parallel-plate capacitor of capacitance 5.0 pF are 2.0 mm apart. What is the area of each plate?

Solution:

$1.1 \times 10^{-3} \mathrm{m}^2$

Exercise:

Problem:

A 60.0-pF vacuum capacitor has a plate area of $0.010 \mathrm{~m}^2$. What is the separation between its plates?

Exercise:

Problem:

A set of parallel plates has a capacitance of $5.0 \mu \mathrm{F}$. How much charge must be added to the plates to increase the potential difference between them by 100 V?

Solution:

500 μC

Exercise:

Problem:

Consider Earth to be a spherical conductor of radius 6400 km and calculate its capacitance.

Exercise:

Problem:

If the capacitance per unit length of a cylindrical capacitor is 20 pF/m, what is the ratio of the radii of the two cylinders?

Solution:

1:16

Exercise:

Problem:

An empty parallel-plate capacitor has a capacitance of $20 \mu \mathrm{F}$. How much charge must leak off its plates before the voltage across them is reduced by 100 V?

Glossary

capacitance
amount of charge stored per unit volt

capacitor
device that stores electrical charge and electrical energy

dielectric
insulating material used to fill the space between two plates

parallel-plate capacitor
system of two identical parallel conducting plates separated by a distance

Capacitors in Series and in Parallel
By the end of this section, you will be able to:

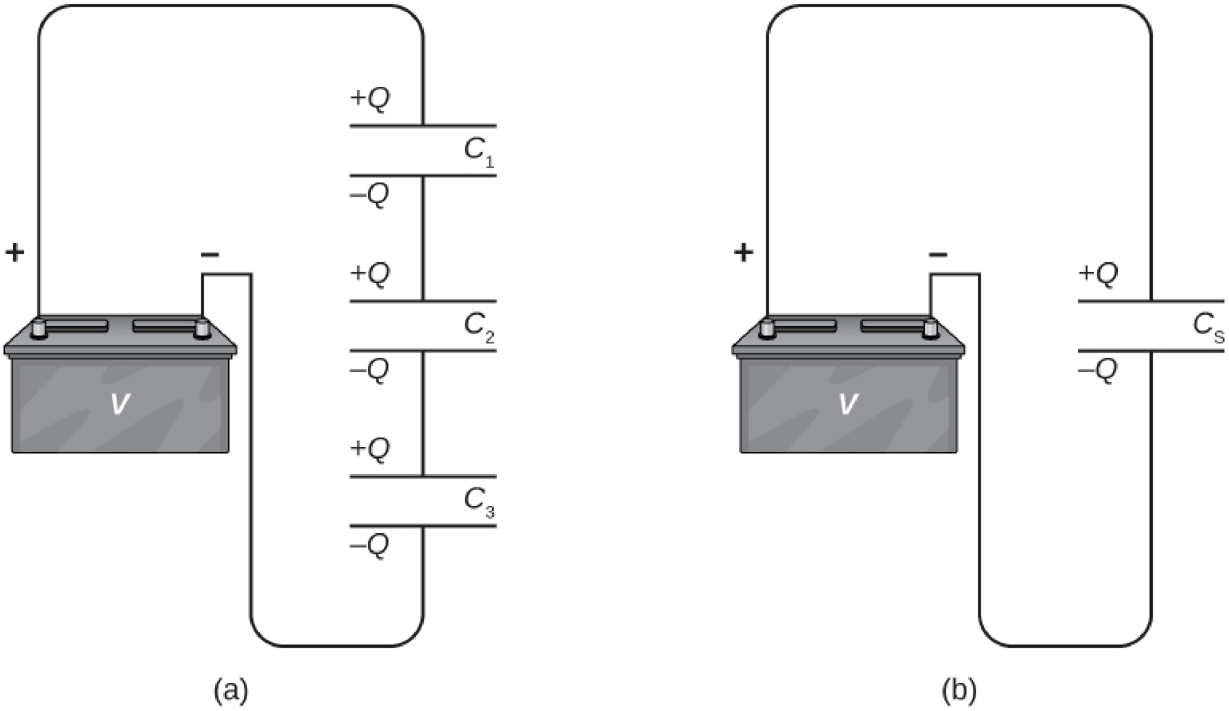
- Explain how to determine the equivalent capacitance of capacitors in series and in parallel combinations
- Compute the potential difference across the plates and the charge on the plates for a capacitor in a network and determine the net capacitance of a network of capacitors

Several capacitors can be connected together to be used in a variety of applications. Multiple connections of capacitors behave as a single equivalent capacitor. The total capacitance of this equivalent single capacitor depends both on the individual capacitors and how they are connected. Capacitors can be arranged in two simple and common types of connections, known as *series* and *parallel*, for which we can easily calculate the total capacitance. These two basic combinations, series and parallel, can also be used as part of more complex connections.

The Series Combination of Capacitors

[link] illustrates a series combination of three capacitors, arranged in a row within the circuit. As for any capacitor, the capacitance of the combination is related to the charge and voltage by using [link]. When this series combination is connected to a battery with voltage V, each of the capacitors acquires an identical charge Q. To explain, first note that the charge on the plate connected to the positive terminal of the battery is $+Q$ and the charge on the plate connected to the negative terminal is $-Q$. Charges are then induced on the other plates so that the sum of the charges on all plates, and the sum of charges on any pair of capacitor plates, is zero. However, the potential drop $V_1 = Q/C_1$ on one capacitor may be different from the potential drop $V_2 = Q/C_2$ on another capacitor, because, generally, the capacitors may have different capacitances. The series combination of two or three capacitors resembles a single capacitor with a smaller capacitance. Generally, any number of capacitors connected in series is equivalent to one capacitor whose capacitance (called the *equivalent capacitance*) is smaller than the smallest of the capacitances in the series combination. Charge on this equivalent capacitor is the same as the charge on any capacitor in a series combination: That is, *all capacitors of a series combination have the same charge*. This occurs due to

the conservation of charge in the circuit. When a charge Q in a series circuit is removed from a plate of the first capacitor (which we denote as $-Q$), it must be placed on a plate of the second capacitor (which we denote as $+Q$), and so on.



(a) Three capacitors are connected in series. The magnitude of the charge on each plate is Q. (b) The network of capacitors in (a) is equivalent to one capacitor that has a smaller capacitance than any of the individual capacitances in (a), and the charge on its plates is Q.

We can find an expression for the total (equivalent) capacitance by considering the voltages across the individual capacitors. The potentials across capacitors 1, 2, and 3 are, respectively, $V_1 = Q/C_1$, $V_2 = Q/C_2$, and $V_3 = Q/C_3$. These potentials must sum up to the voltage of the battery, giving the following potential balance:

Equation:

$$V = V_1 + V_2 + V_3.$$

Potential V is measured across an equivalent capacitor that holds charge Q and has an equivalent capacitance C_{S}. Entering the expressions for V_1, V_2, and V_3, we get

Equation:

$$\frac{Q}{C_{\text{S}}} = \frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3}.$$

Canceling the charge Q, we obtain an expression containing the equivalent capacitance, C_{S}, of three capacitors connected in series:

Equation:

$$\frac{1}{C_{\text{S}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}.$$

This expression can be generalized to any number of capacitors in a series network.

Note:

Series Combination

For capacitors connected in a **series combination**, the reciprocal of the equivalent capacitance is the sum of reciprocals of individual capacitances:

Equation:

$$\frac{1}{C_{\text{S}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots.$$

Example:

Equivalent Capacitance of a Series Network

Find the total capacitance for three capacitors connected in series, given their individual capacitances are $1.000\ \mu\text{F}$, $5.000\ \mu\text{F}$, and $8.000\ \mu\text{F}$.

Strategy

Because there are only three capacitors in this network, we can find the equivalent capacitance by using [link] with three terms.

Solution

We enter the given capacitances into [link]:

Equation:

$$\begin{aligned}\frac{1}{C_\text{S}} &= \frac{1}{C_1}+\frac{1}{C_2}+\frac{1}{C_3}\\ &= \frac{1}{1.000\ \mu\text{F}}+\frac{1}{5.000\ \mu\text{F}}+\frac{1}{8.000\ \mu\text{F}}\\ \frac{1}{C_\text{S}} &= \frac{1.325}{\mu\text{F}}.\end{aligned}$$

Now we invert this result and obtain $C_\text{S} = \frac{\mu\text{F}}{1.325} = 0.755\ \mu\text{F}.$

Significance

Note that in a series network of capacitors, the equivalent capacitance is always less than the smallest individual capacitance in the network.

The Parallel Combination of Capacitors

A parallel combination of three capacitors, with one plate of each capacitor connected to one side of the circuit and the other plate connected to the other side, is illustrated in [link](a). Since the capacitors are connected in parallel, *they all have the same voltage V across their plates*. However, each capacitor in the parallel network may store a different charge. To find the equivalent capacitance C_P of the parallel network, we note that the total charge Q stored by the network is the sum of all the individual charges:

Equation:

$$Q = Q_1 + Q_2 + Q_3.$$

On the left-hand side of this equation, we use the relation $Q = C_\text{P}V$, which holds for the entire network. On the right-hand side of the equation, we use the relations $Q_1 = C_1V, Q_2 = C_2V$, and $Q_3 = C_3V$ for the three capacitors in the network. In this way we obtain

Equation:

$$C_{\text{P}}V = C_1V + C_2V + C_3V.$$

This equation, when simplified, is the expression for the equivalent capacitance of the parallel network of three capacitors:

Equation:

$$C_{\text{P}} = C_1 + C_2 + C_3.$$

This expression is easily generalized to any number of capacitors connected in parallel in the network.

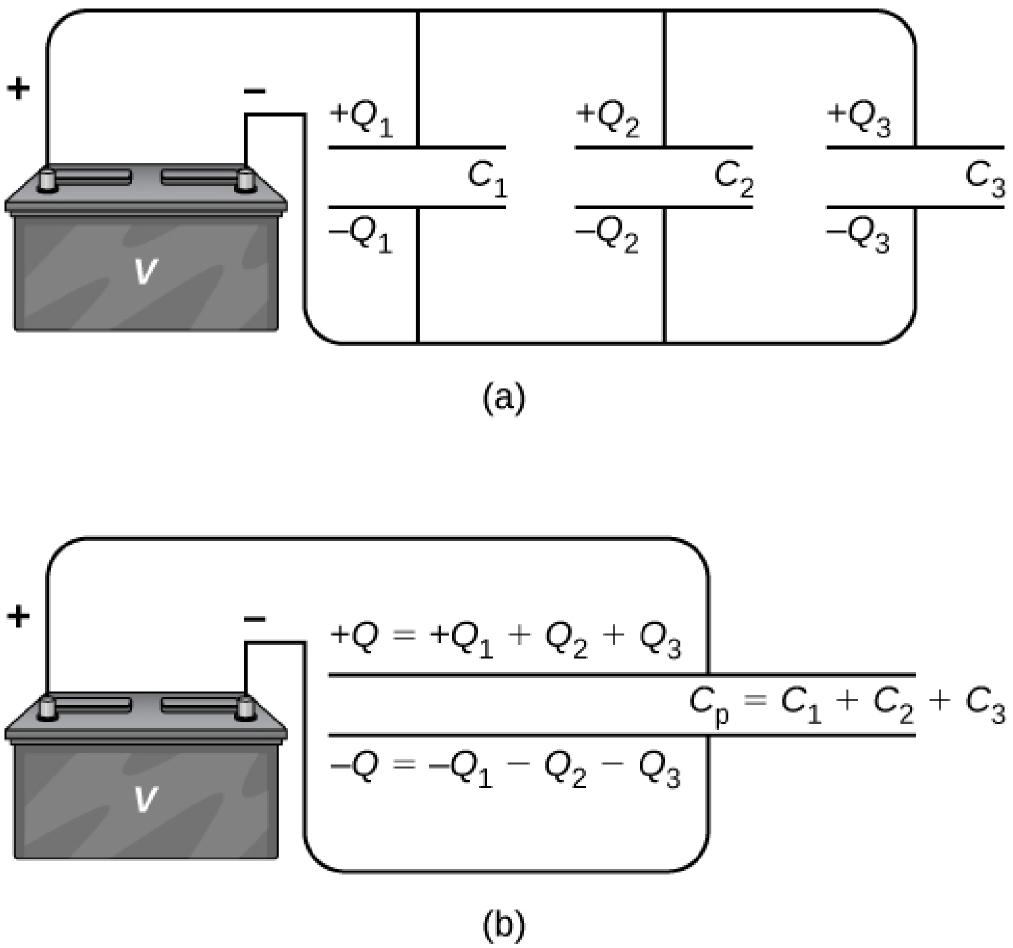
Note:

Parallel Combination

For capacitors connected in a **parallel combination**, the equivalent (net) capacitance is the sum of all individual capacitances in the network,

Equation:

$$C_{\text{P}} = C_1 + C_2 + C_3 + \cdots.$$



(a) Three capacitors are connected in parallel. Each capacitor is connected directly to the battery. (b) The charge on the equivalent capacitor is the sum of the charges on the individual capacitors.

Example:

Equivalent Capacitance of a Parallel Network

Find the net capacitance for three capacitors connected in parallel, given their individual capacitances are $1.0\,\mu\text{F}, 5.0\,\mu\text{F},$ and $8.0\,\mu\text{F}$.

Strategy

Because there are only three capacitors in this network, we can find the equivalent capacitance by using [link] with three terms.

Solution

Entering the given capacitances into [link] yields

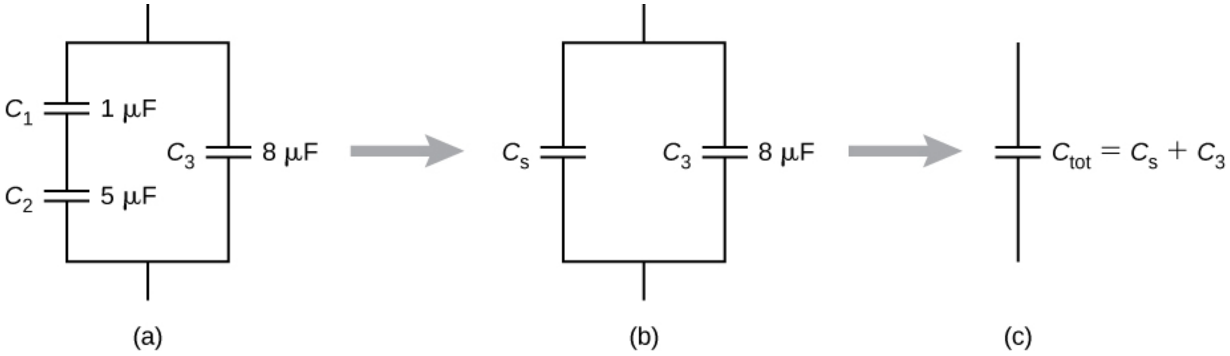
Equation:

$$\begin{aligned} C_P &= C_1 + C_2 + C_3 = 1.0\,\mu\text{F} + 5.0\,\mu\text{F} + 8.0\,\mu\text{F} \\ C_P &= 14.0\,\mu\text{F}. \end{aligned}$$

Significance

Note that in a parallel network of capacitors, the equivalent capacitance is always larger than any of the individual capacitances in the network.

Capacitor networks are usually some combination of series and parallel connections, as shown in [link]. To find the net capacitance of such combinations, we identify parts that contain only series or only parallel connections, and find their equivalent capacitances. We repeat this process until we can determine the equivalent capacitance of the entire network. The following example illustrates this process.



(a) This circuit contains both series and parallel connections of capacitors. (b) C_1 and C_2 are in series; their equivalent capacitance is C_S. (c) The equivalent capacitance C_S is connected in parallel with C_3. Thus, the equivalent capacitance of the entire network is the sum of C_S and C_3.

Example:

Equivalent Capacitance of a Network

Find the total capacitance of the combination of capacitors shown in [link]. Assume the capacitances are known to three decimal places $(C_1 = 1.000\ \mu\text{F}, C_2 = 5.000\ \mu\text{F}, C_3 = 8.000\ \mu\text{F})$. Round your answer to three decimal places.

Strategy

We first identify which capacitors are in series and which are in parallel. Capacitors C_1 and C_2 are in series. Their combination, labeled C_S, is in parallel with C_3.

Solution

Since C_1 and C_2 are in series, their equivalent capacitance C_S is obtained with [link]:

Equation:

$$\frac{1}{C_\text{S}} = \frac{1}{C_1} + \frac{1}{C_2} = \frac{1}{1.000\ \mu\text{F}} + \frac{1}{5.000\ \mu\text{F}} = \frac{1.200}{\mu\text{F}} \Rightarrow C_\text{S} = 0.833\ \mu\text{F}.$$

Capacitance C_S is connected in parallel with the third capacitance C_3, so we use [link] to find the equivalent capacitance C of the entire network:

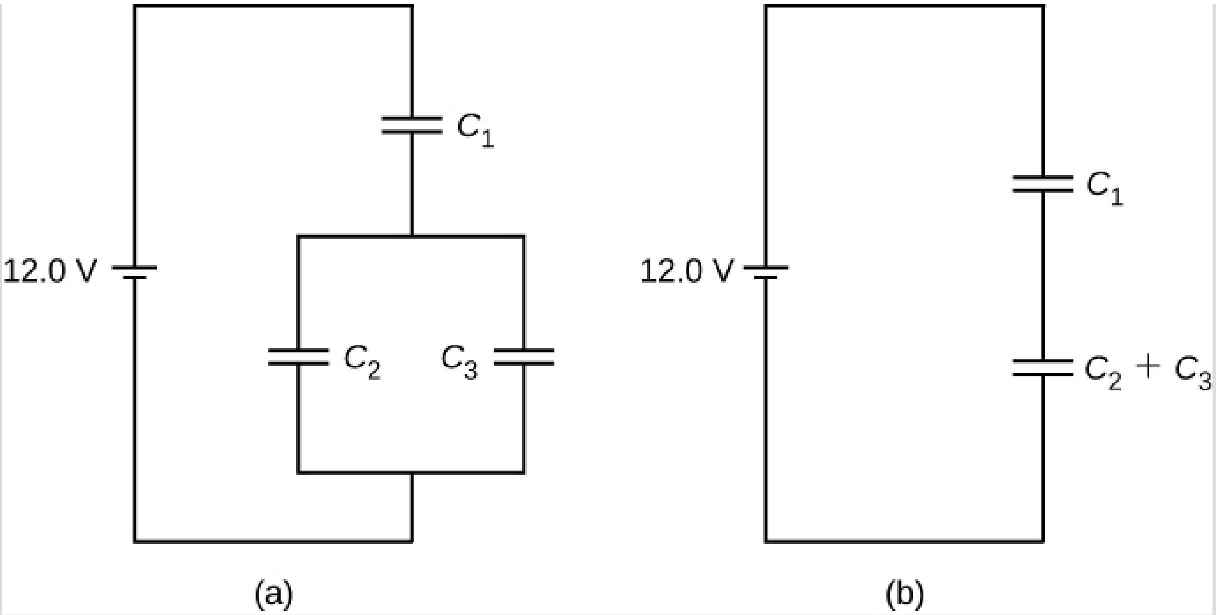
Equation:

$$C = C_\text{S} + C_3 = 0.833\ \mu\text{F} + 8.000\ \mu\text{F} = 8.833\ \mu\text{F}.$$

Example:

Network of Capacitors

Determine the net capacitance C of the capacitor combination shown in [link] when the capacitances are $C_1 = 12.0\ \mu\text{F}, C_2 = 2.0\ \mu\text{F}$, and $C_3 = 4.0\ \mu\text{F}$. When a 12.0-V potential difference is maintained across the combination, find the charge and the voltage across each capacitor.



(a) A capacitor combination. (b) An equivalent two-capacitor combination.

Strategy

We first compute the net capacitance C_{23} of the parallel connection C_2 and C_3. Then C is the net capacitance of the series connection C_1 and C_{23}. We use the relation $C = Q/V$ to find the charges Q_1,Q_2, and Q_3, and the voltages V_1, V_2, and V_3, across capacitors 1, 2, and 3, respectively.

Solution

The equivalent capacitance for C_2 and C_3 is

Equation:

$$C_{23} = C_2 + C_3 = 2.0\,\mu\text{F} + 4.0\,\mu\text{F} = 6.0\,\mu\text{F}.$$

The entire three-capacitor combination is equivalent to two capacitors in series,

Equation:

$$\frac{1}{C} = \frac{1}{12.0\,\mu\text{F}} + \frac{1}{6.0\,\mu\text{F}} = \frac{1}{4.0\,\mu\text{F}} \Rightarrow C = 4.0\,\mu\text{F}.$$

Consider the equivalent two-capacitor combination in [link](b). Since the capacitors are in series, they have the same charge, $Q_1 = Q_{23}$. Also, the capacitors share the 12.0-V potential difference, so

Equation:

$$12.0\ \text{V} = V_1 + V_{23} = \frac{Q_1}{C_1} + \frac{Q_{23}}{C_{23}} = \frac{Q_1}{12.0\ \mu\text{F}} + \frac{Q_1}{6.0\ \mu\text{F}} \Rightarrow Q_1 = 48.0\ \mu\text{C}.$$

Now the potential difference across capacitor 1 is

Equation:

$$V_1 = \frac{Q_1}{C_1} = \frac{48.0\ \mu\text{C}}{12.0\ \mu\text{F}} = 4.0\ \text{V}.$$

Because capacitors 2 and 3 are connected in parallel, they are at the same potential difference:

Equation:

$$V_2 = V_3 = 12.0\ \text{V} - 4.0\ \text{V} = 8.0\ \text{V}.$$

Hence, the charges on these two capacitors are, respectively,

Equation:

$$\begin{aligned} Q_2 &= C_2V_2 = (2.0\ \mu\text{F})(8.0\ \text{V}) = 16.0\ \mu\text{C}, \\ Q_3 &= C_3V_3 = (4.0\ \mu\text{F})(8.0\ \text{V}) = 32.0\ \mu\text{C}. \end{aligned}$$

Significance

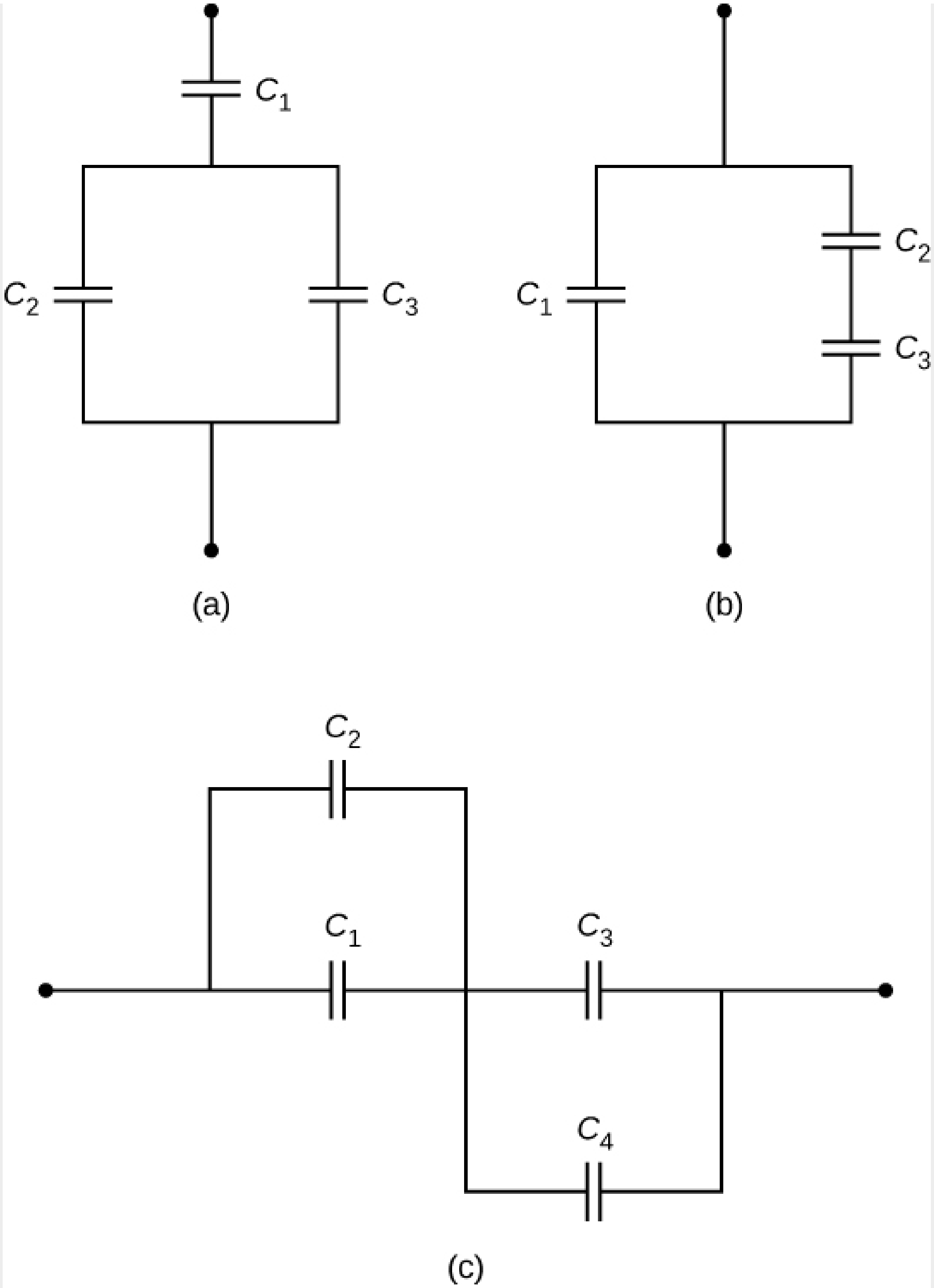
As expected, the net charge on the parallel combination of C_2 and C_3 is $Q_{23} = Q_2 + Q_3 = 48.0\ \mu\text{C}$.

Note:

Exercise:

Problem:

Check Your Understanding Determine the net capacitance C of each network of capacitors shown below. Assume that $C_1 = 1.0\ \text{pF}$, $C_2 = 2.0\ \text{pF}$, $C_3 = 4.0\ \text{pF}$, and $C_4 = 5.0\ \text{pF}$. Find the charge on each capacitor, assuming there is a potential difference of 12.0 V across each network.


C_1
C_2
C_3
(a)
C_2
C_1
C_3
(b)
C_2
C_1
C_3
C_4
(c)

Solution:

a. $C = 0.86\,\text{pF}, Q_1 = 10\,\text{pC}, Q_2 = 3.4\,\text{pC}, Q_3 = 6.8\,\text{pC};$
b. $C = 2.3\,\text{pF}, Q_1 = 12\,\text{pC}, Q_2 = Q_3 = 16\,\text{pC};$

c. $C = 2.3\ \mathrm{pF}, Q_1 = 9.0\ \mathrm{pC}, Q_2 = 18\ \mathrm{pC}, Q_3 = 12\ \mathrm{pC}, Q_4 = 15\ \mathrm{pC}$

Summary

- When several capacitors are connected in a series combination, the reciprocal of the equivalent capacitance is the sum of the reciprocals of the individual capacitances.
- When several capacitors are connected in a parallel combination, the equivalent capacitance is the sum of the individual capacitances.
- When a network of capacitors contains a combination of series and parallel connections, we identify the series and parallel networks, and compute their equivalent capacitances step by step until the entire network becomes reduced to one equivalent capacitance.

Conceptual Questions

Exercise:

Problem:

If you wish to store a large amount of charge in a capacitor bank, would you connect capacitors in series or in parallel? Explain.

Exercise:

Problem:

What is the maximum capacitance you can get by connecting three $1.0\text{-}\mu\mathrm{F}$ capacitors? What is the minimum capacitance?

Solution:

$3.0\ \mu\mathrm{F}, 0.33\ \mu\mathrm{F}$

Problems

Exercise:

Problem:

A 4.00-pF is connected in series with an 8.00-pF capacitor and a 400-V potential difference is applied across the pair. (a) What is the charge on each capacitor? (b) What is the voltage across each capacitor?

Solution:

a. 1.07 nC; b. 267 V, 133 V

Exercise:

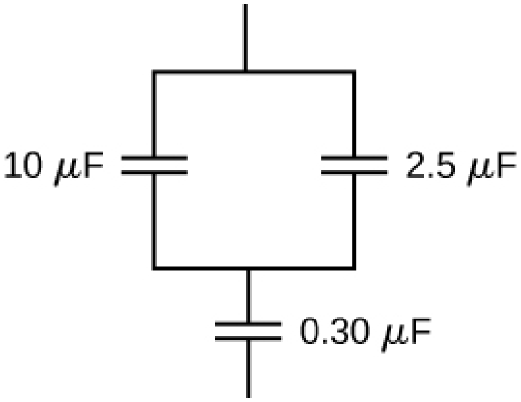
Problem:

Three capacitors, with capacitances of $C_1 = 2.0\ \mu\text{F}$,$C_2 = 3.0\ \mu\text{F}$, and $C_3 = 6.0\ \mu\text{F}$, respectively, are connected in parallel. A 500-V potential difference is applied across the combination. Determine the voltage across each capacitor and the charge on each capacitor.

Exercise:

Problem:

Find the total capacitance of this combination of series and parallel capacitors shown below.



Solution:

$0.29\ \mu\text{F}$

Exercise:

Problem:

Suppose you need a capacitor bank with a total capacitance of 0.750 F but you have only 1.50-mF capacitors at your disposal. What is the smallest number of capacitors you could connect together to achieve your goal, and how would you connect them?

Solution:

500 capacitors; connected in parallel

Exercise:

Problem:

What total capacitances can you make by connecting a $5.00\text{-}\mu\text{F}$ and a $8.00\text{-}\mu\text{F}$ capacitor?

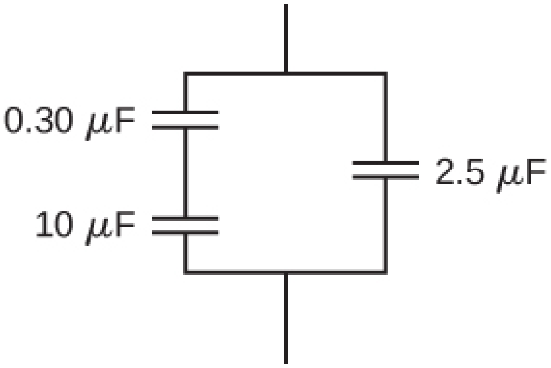
Solution:

$3.08\ \mu\text{F}$ (series) and $13.0\ \mu\text{F}$ (parallel)

Exercise:

Problem:

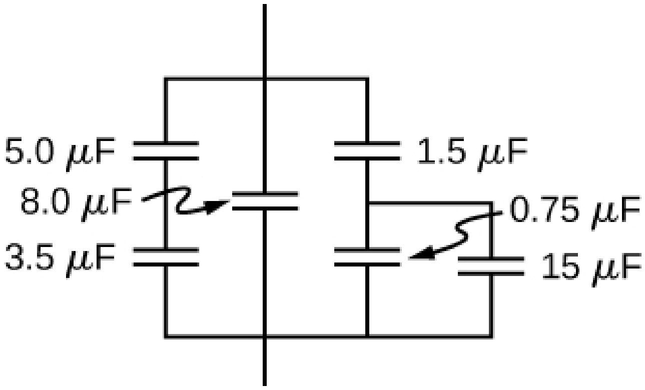
Find the equivalent capacitance of the combination of series and parallel capacitors shown below.



Exercise:

Problem:

Find the net capacitance of the combination of series and parallel capacitors shown below.



Solution:

$11.4\ \mu\text{F}$

Exercise:

Problem:

A 40-pF capacitor is charged to a potential difference of 500 V. Its terminals are then connected to those of an uncharged 10-pF capacitor. Calculate: (a) the original charge on the 40-pF capacitor; (b) the charge on each capacitor after the connection is made; and (c) the potential difference across the plates of each capacitor after the connection.

Exercise:

Problem:

A $2.0\text{-}\mu\text{F}$ capacitor and a $4.0\text{-}\mu\text{F}$ capacitor are connected in series across a 1.0-kV potential. The charged capacitors are then disconnected from the source and connected to each other with terminals of like sign together. Find the charge on each capacitor and the voltage across each capacitor.

Solution:

0.89 mC; 1.78 mC; 444 V

Glossary

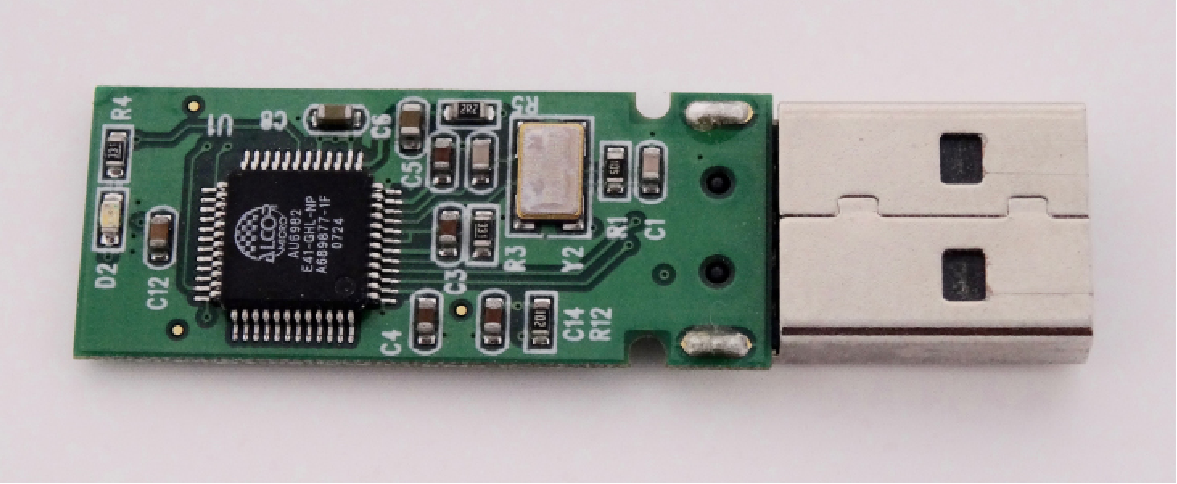
parallel combination	components in a circuit arranged with one side of each component connected to one side of the circuit and the other sides of the components connected to the other side of the circuit
series combination	components in a circuit arranged in a row one after the other in a circuit

Energy Stored in a Capacitor

By the end of this section, you will be able to:

- Explain how energy is stored in a capacitor
- Use energy relations to determine the energy stored in a capacitor network

Most of us have seen dramatizations of medical personnel using a defibrillator to pass an electrical current through a patient's heart to get it to beat normally. Often realistic in detail, the person applying the shock directs another person to "make it 400 joules this time." The energy delivered by the defibrillator is stored in a capacitor and can be adjusted to fit the situation. SI units of joules are often employed. Less dramatic is the use of capacitors in microelectronics to supply energy when batteries are charged ([link]). Capacitors are also used to supply energy for flash lamps on cameras.



The capacitors on the circuit board for an electronic device follow a labeling convention that identifies each one with a code that begins with the letter "C." (credit: Windell Oskay)

The energy U_C stored in a capacitor is electrostatic potential energy and is thus related to the charge Q and voltage V between the capacitor plates. A charged capacitor stores energy in the electrical field between its plates. As the capacitor is being charged, the electrical field builds up. When a charged capacitor is disconnected from a battery, its energy remains in the field in the space between its plates.

To gain insight into how this energy may be expressed (in terms of Q and V), consider a charged, empty, parallel-plate capacitor; that is, a capacitor without a dielectric but with a vacuum between its plates. The space between its plates has a volume Ad, and it is filled with a uniform electrostatic field E. The total energy U_C of the capacitor is contained within this space. The **energy density** u_E in this space is simply U_C divided by the volume Ad. If we know the energy density, the energy can be found as $U_C = u_E(Ad)$. We will learn in Electromagnetic Waves (after completing the study of Maxwell's equations) that the energy density u_E in a region of free space occupied by an electrical field E depends only on the magnitude of the field and is

Note:
Equation:

$$u_E = \frac{1}{2}\varepsilon_0 E^2.$$

If we multiply the energy density by the volume between the plates, we obtain the amount of energy stored between the plates of a parallel-plate capacitor: $U_C = u_E(Ad) = \frac{1}{2}\varepsilon_0 E^2 Ad = \frac{1}{2}\varepsilon_0 \frac{V^2}{d^2} Ad = \frac{1}{2}V^2\varepsilon_0 \frac{A}{d} = \frac{1}{2}V^2 C.$

In this derivation, we used the fact that the electrical field between the plates is uniform so that $E = V/d$ and $C = \varepsilon_0 A/d$. Because $C = Q/V$, we can express this result in other equivalent forms:

Note:
Equation:

$$U_C = \frac{1}{2}V^2 C = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}QV.$$

The expression in [link] for the energy stored in a parallel-plate capacitor is generally valid for all types of capacitors. To see this, consider any uncharged capacitor (not necessarily a parallel-plate type). At some instant, we connect it across a battery, giving it a potential difference $V = q/C$ between its plates. Initially, the charge on the plates is $Q = 0$. As the capacitor is being charged, the charge gradually builds up on its plates, and after some time, it reaches the value Q. To move an infinitesimal charge dq from the negative plate to the positive plate (from a lower to a higher potential), the amount of work dW that must be done on dq is $dW = Vdq = \frac{q}{C}dq.$

This work becomes the energy stored in the electrical field of the capacitor. In order to charge the capacitor to a charge Q, the total work required is
Equation:

$$W = \int_0^{W(Q)} dW = \int_0^Q \frac{q}{C} dq = \frac{1}{2} \frac{Q^2}{C}.$$

Since the geometry of the capacitor has not been specified, this equation holds for any type of capacitor. The total work W needed to charge a capacitor is the electrical potential energy U_C stored in it, or $U_C = W$. When the charge is expressed in coulombs, potential is expressed in volts, and the capacitance is expressed in farads, this relation gives the energy in joules.

Knowing that the energy stored in a capacitor is $U_C = Q^2/(2C)$, we can now find the energy density u_E stored in a vacuum between the plates of a charged parallel-plate capacitor. We just have to divide U_C by the volume *Ad* of space between its plates and take into account that for a parallel-plate capacitor, we have $E = \sigma/\varepsilon_0$ and $C = \varepsilon_0 A/d$. Therefore, we obtain

Equation:

$$u_E = \frac{U_C}{Ad} = \frac{1}{2} \frac{Q^2}{C} \frac{1}{Ad} = \frac{1}{2} \frac{Q^2}{\varepsilon_0 A/d} \frac{1}{Ad} = \frac{1}{2} \frac{1}{\varepsilon_0} \left(\frac{Q}{A}\right)^2 = \frac{\sigma^2}{2\varepsilon_0} = \frac{(E\varepsilon_0)^2}{2\varepsilon_0} = \frac{\varepsilon_0}{2} E^2.$$

We see that this expression for the density of energy stored in a parallel-plate capacitor is in accordance with the general relation expressed in [link]. We could repeat this calculation for either a spherical capacitor or a cylindrical capacitor—or other capacitors—and in all cases, we would end up with the general relation given by [link].

Example:

Energy Stored in a Capacitor

Calculate the energy stored in the capacitor network in [link](a) when the capacitors are fully charged and when the capacitances are $C_1 = 12.0\,\mu\text{F}$, $C_2 = 2.0\,\mu\text{F}$, and $C_3 = 4.0\,\mu\text{F}$, respectively.

Strategy

We use [link] to find the energy U_1, U_2, and U_3 stored in capacitors 1, 2, and 3, respectively. The total energy is the sum of all these energies.

Solution

We identify $C_1 = 12.0\,\mu\text{F}$ and $V_1 = 4.0\,\text{V}$, $C_2 = 2.0\,\mu\text{F}$ and $V_2 = 8.0\,\text{V}$, $C_3 = 4.0\,\mu\text{F}$ and $V_3 = 8.0\,\text{V}$. The energies stored in these capacitors are

Equation:

$$\begin{aligned} U_1 &= \tfrac{1}{2} C_1 V_1^2 = \tfrac{1}{2}(12.0\,\mu\text{F})(4.0\,\text{V})^2 = 96\,\mu\text{J}, \\ U_2 &= \tfrac{1}{2} C_2 V_2^2 = \tfrac{1}{2}(2.0\,\mu\text{F})(8.0\,\text{V})^2 = 64\,\mu\text{J}, \\ U_3 &= \tfrac{1}{2} C_3 V_3^2 = \tfrac{1}{2}(4.0\,\mu\text{F})(8.0\,\text{V})^2 = 130\,\mu\text{J}. \end{aligned}$$

The total energy stored in this network is

Equation:

$$U_C = U_1 + U_2 + U_3 = 96\,\mu\text{J} + 64\,\mu\text{J} + 130\,\mu\text{J} = 0.29\,\text{mJ}.$$

Significance

We can verify this result by calculating the energy stored in the single 4.0-μF capacitor, which is found to be equivalent to the entire network. The voltage across the network is 12.0 V. The total energy obtained in this way agrees with our previously obtained result, $U_C = \frac{1}{2}CV^2 = \frac{1}{2}(4.0\,\mu\text{F})(12.0\,\text{V})^2 = 0.29\,\text{mJ}.$

Note:

Exercise:

Problem:

Check Your Understanding The potential difference across a 5.0-pF capacitor is 0.40 V. (a) What is the energy stored in this capacitor? (b) The potential difference is now increased to 1.20 V. By what factor is the stored energy increased?

Solution:

a.4.0×10^{-13}J; b. 9 times

In a cardiac emergency, a portable electronic device known as an automated external defibrillator (AED) can be a lifesaver. A **defibrillator** ([link]) delivers a large charge in a short burst, or a shock, to a person's heart to correct abnormal heart rhythm (an arrhythmia). A heart attack can arise from the onset of fast, irregular beating of the heart—called cardiac or ventricular fibrillation. Applying a large shock of electrical energy can terminate the arrhythmia and allow the body's natural pacemaker to resume its normal rhythm. Today, it is common for ambulances to carry AEDs. AEDs are also found in many public places. These are designed to be used by lay persons. The device automatically diagnoses the patient's heart rhythm and then applies the shock with appropriate energy and waveform. CPR (cardiopulmonary resuscitation) is recommended in many cases before using a defibrillator.

Automated external defibrillators are found in many public places. These portable units provide verbal instructions for use in the important first few minutes for a person suffering a cardiac attack. (credit: Owain Davies)

Example:

Capacitance of a Heart Defibrillator

A heart defibrillator delivers 4.00×10^2J of energy by discharging a capacitor initially at 1.00×10^4 V. What is its capacitance?

Strategy

We are given U_C and V, and we are asked to find the capacitance C. We solve [link] for C and substitute.

Solution

Solving this expression for C and entering the given values yields

$$C = 2\frac{U_C}{V^2} = 2\frac{4.00 \times 10^2 \text{ J}}{(1.00 \times 10^4 \text{ V})^2} = 8.00\,\mu\text{F}.$$

Summary

- Capacitors are used to supply energy to a variety of devices, including defibrillators, microelectronics such as calculators, and flash lamps.
- The energy stored in a capacitor is the work required to charge the capacitor, beginning with no charge on its plates. The energy is stored in the electrical field in the space between the capacitor plates. It depends on the amount of electrical charge on the plates and on the potential difference between the plates.
- The energy stored in a capacitor network is the sum of the energies stored on individual capacitors in the network. It can be computed as the energy stored in the equivalent capacitor of the network.

Conceptual Questions

Exercise:

Problem:

If you wish to store a large amount of energy in a capacitor bank, would you connect capacitors in series or parallel? Explain.

Problems

Exercise:

Problem:

How much energy is stored in an $8.00\text{-}\mu\text{F}$ capacitor whose plates are at a potential difference of 6.00 V?

Exercise:

Problem:

A capacitor has a charge of $2.5\ \mu\text{C}$ when connected to a 6.0-V battery. How much energy is stored in this capacitor?

Solution:

$7.5\ \mu\text{J}$

Exercise:

Problem:

How much energy is stored in the electrical field of a metal sphere of radius 2.0 m that is kept at a 10.0-V potential?

Exercise:

Problem:

(a) What is the energy stored in the $10.0\text{-}\mu\text{F}$ capacitor of a heart defibrillator charged to 9.00×10^3 V ? (b) Find the amount of the stored charge.

Solution:

a. 405 J; b. 90.0 mC

Exercise:

Problem:

In open-heart surgery, a much smaller amount of energy will defibrillate the heart. (a) What voltage is applied to the $8.00\text{-}\mu\text{F}$ capacitor of a heart defibrillator that stores 40.0 J of energy? (b) Find the amount of the stored charge.

Exercise:

Problem:

A $165\text{-}\mu\text{F}$ capacitor is used in conjunction with a dc motor. How much energy is stored in it when 119 V is applied?

Solution:

1.17 J

Exercise:

Problem:

Suppose you have a 9.00-V battery, a $2.00\text{-}\mu\text{F}$ capacitor, and a $7.40\text{-}\mu\text{F}$ capacitor. (a) Find the charge and energy stored if the capacitors are connected to the battery in series. (b) Do the same for a parallel connection.

Exercise:

Problem:

An anxious physicist worries that the two metal shelves of a wood frame bookcase might obtain a high voltage if charged by static electricity, perhaps produced by friction. (a) What is the capacitance of the empty shelves if they have area $1.00 \times 10^2\ \text{m}^2$ and are 0.200 m apart? (b) What is the voltage between them if opposite charges of magnitude 2.00 nC are placed on them? (c) To show that this voltage poses a small hazard, calculate the energy stored. (d) The actual shelves have an area 100 times smaller than these hypothetical shelves. Are his fears justified?

Solution:

a. 4.43×10^{-9} F; b. 0.453 V; c. 4.53×10^{-10} J; d. no

Exercise:

Problem:

A parallel-plate capacitor is made of two square plates 25 cm on a side and 1.0 mm apart. The capacitor is connected to a 50.0-V battery. With the battery still connected, the plates are pulled apart to a separation of 2.00 mm. What are the energies stored in the capacitor before and after the plates are pulled farther apart? Why does the energy decrease even though work is done in separating the plates?

Exercise:

Problem:

Suppose that the capacitance of a variable capacitor can be manually changed from 100 pF to 800 pF by turning a dial, connected to one set of plates by a shaft, from 0° to 180°. With the dial set at 180° (corresponding to $C = 800\,\text{pF}$), the capacitor is connected to a 500-V source. After charging, the capacitor is disconnected from the source, and the dial is turned to 0°. If friction is negligible, how much work is required to turn the dial from 180° to 0° ?

Solution:

0.7 mJ

Glossary

energy density
: energy stored in a capacitor divided by the volume between the plates

Capacitor with a Dielectric

By the end of this section, you will be able to:

- Describe the effects a dielectric in a capacitor has on capacitance and other properties
- Calculate the capacitance of a capacitor containing a dielectric

As we discussed earlier, an insulating material placed between the plates of a capacitor is called a dielectric. Inserting a dielectric between the plates of a capacitor affects its capacitance. To see why, let's consider an experiment described in [link]. Initially, a capacitor with capacitance C_0 when there is air between its plates is charged by a battery to voltage V_0. When the capacitor is fully charged, the battery is disconnected. A charge Q_0 then resides on the plates, and the potential difference between the plates is measured to be V_0. Now, suppose we insert a dielectric that *totally* fills the gap between the plates. If we monitor the voltage, we find that the voltmeter reading has dropped to a *smaller* value V. We write this new voltage value as a fraction of the original voltage V_0, with a positive number κ, $\kappa > 1$:

Equation:

$$V = \frac{1}{\kappa} V_0.$$

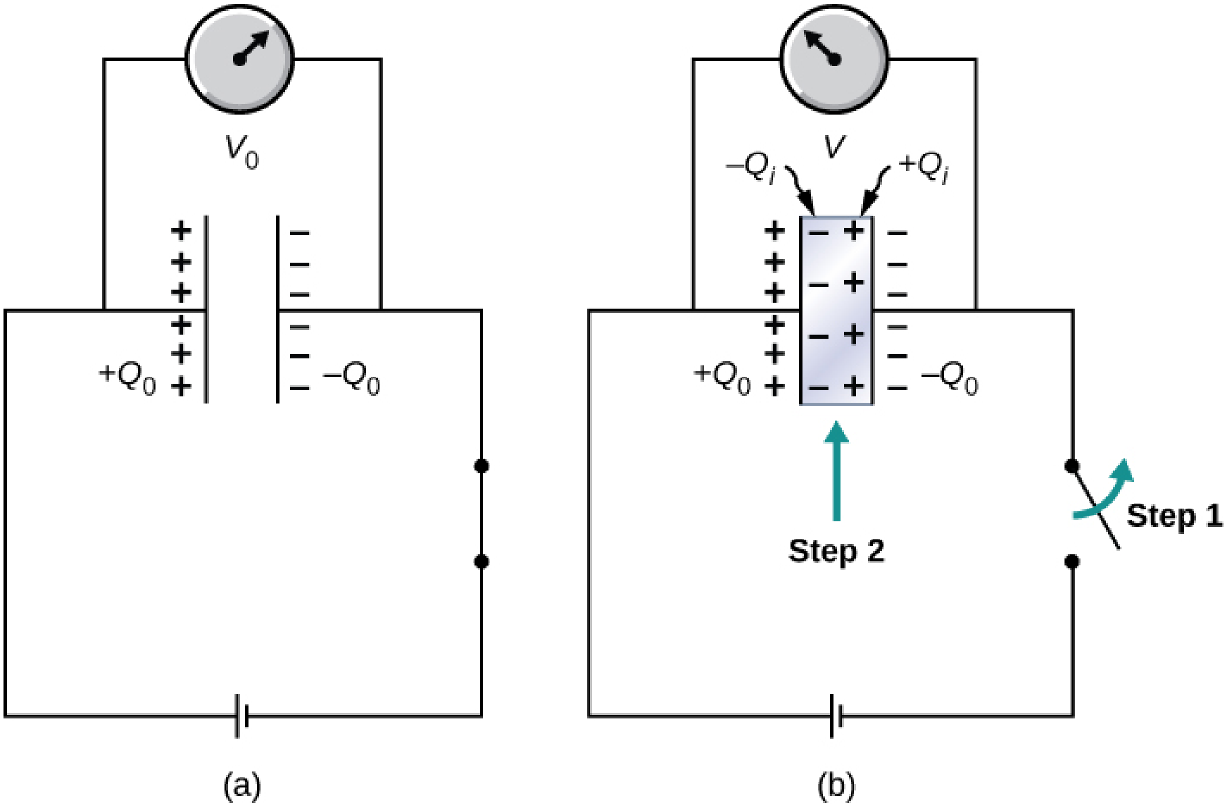
The constant κ in this equation is called the **dielectric constant** of the material between the plates, and its value is characteristic for the material. A detailed explanation for why the dielectric reduces the voltage is given in the next section. Different materials have different dielectric constants (a table of values for typical materials is provided in the next section). Once the battery becomes disconnected, there is no path for a charge to flow to the battery from the capacitor plates. Hence, the insertion of the dielectric has no effect on the charge on the plate, which remains at a value of Q_0. Therefore, we find that the capacitance of the capacitor with a dielectric is

Note:

Equation:

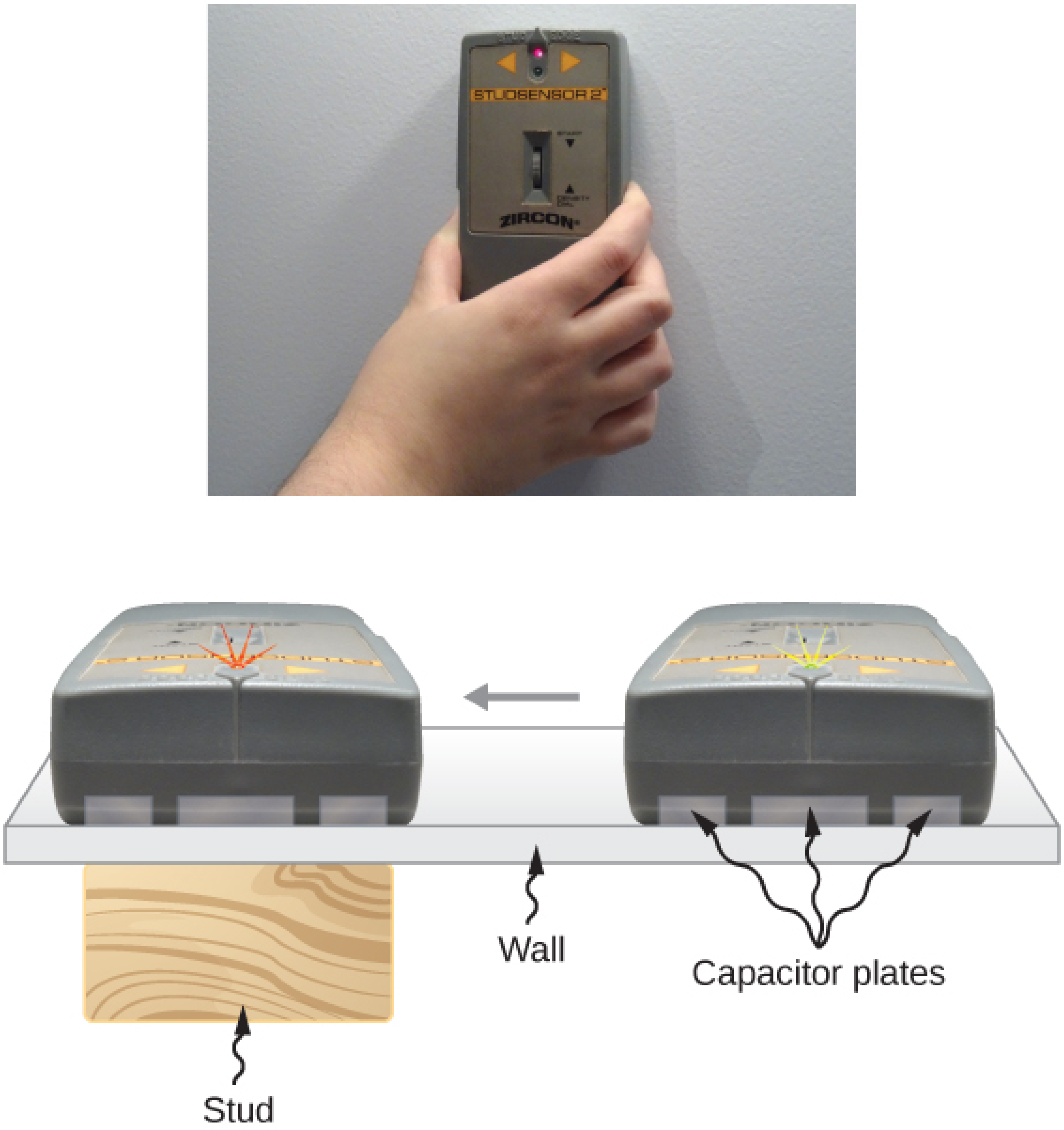
$$C = \frac{Q_0}{V} = \frac{Q_0}{V_0/\kappa} = \kappa\frac{Q_0}{V_0} = \kappa C_0.$$

This equation tells us that the *capacitance* C_0 *of an empty (vacuum) capacitor can be increased by a factor of* κ *when we insert a dielectric material to completely fill the space between its plates*. Note that [link] can also be used for an empty capacitor by setting $\kappa = 1$. In other words, we can say that the dielectric constant of the vacuum is 1, which is a reference value.



(a) When fully charged, a vacuum capacitor has a voltage V_0 and charge Q_0 (the charges remain on plate's inner surfaces; the schematic indicates the sign of charge on each plate). (b) In step 1, the battery is disconnected. Then, in step 2, a dielectric (that is electrically neutral) is inserted into the charged capacitor. When the voltage across the capacitor is now measured, it is found that the voltage value has decreased to $V = V_0/\kappa$. The schematic indicates the sign of the induced charge that is now present on the surfaces of the dielectric material between the plates.

The principle expressed by [link] is widely used in the construction industry ([link]). Metal plates in an electronic stud finder act effectively as a capacitor. You place a stud finder with its flat side on the wall and move it continually in the horizontal direction. When the finder moves over a wooden stud, the capacitance of its plates changes, because wood has a different dielectric constant than a gypsum wall. This change triggers a signal in a circuit, and thus the stud is detected.


STUDSENSOR 2
ZIRCON
Wall
Capacitor plates
Stud

An electronic stud finder is used to detect wooden studs behind drywall. (credit top: modification of work by Jane Whitney)

The electrical energy stored by a capacitor is also affected by the presence of a dielectric. When the energy stored in an empty capacitor is U_0, the energy U stored in a capacitor with a dielectric is smaller by a factor of κ,

Note:
Equation:

$$U = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}\frac{Q_0^2}{\kappa C_0} = \frac{1}{\kappa}U_0.$$

As a dielectric material sample is brought near an empty charged capacitor, the sample reacts to the electrical field of the charges on the capacitor plates. Just as we learned in Electric Charges and Fields on electrostatics, there will be the induced charges on the surface of the sample; however, they are not free charges like in a conductor, because a perfect insulator does not have freely moving charges. These induced charges on the dielectric surface are of an opposite sign to the free charges on the plates of the capacitor, and so they are attracted by the free charges on the plates. Consequently, the dielectric is “pulled” into the gap, and the work to polarize the dielectric material between the plates is done at the expense of the stored electrical energy, which is reduced, in accordance with [link].

Example:
Inserting a Dielectric into an Isolated Capacitor
An empty 20.0-pF capacitor is charged to a potential difference of 40.0 V. The charging battery is then disconnected, and a piece of Teflon™ with a dielectric constant of 2.1 is inserted to completely fill the space between the capacitor plates (see [link]). What are the values of (a) the capacitance, (b) the charge of the plate, (c) the potential difference between the plates, and (d) the energy stored in the capacitor with and without dielectric?
Strategy
We identify the original capacitance $C_0 = 20.0\text{ pF}$ and the original potential difference $V_0 = 40.0\text{ V}$ between the plates. We combine [link] with other relations involving capacitance and substitute.
Solution

a. The capacitance increases to

Equation:

$$C = \kappa C_0 = 2.1(20.0\,\text{pF}) = 42.0\,\text{pF}.$$

b. Without dielectric, the charge on the plates is

Equation:

$$Q_0 = C_0V_0 = (20.0\,\text{pF})(40.0\,\text{V}) = 0.8\,\text{nC}.$$

Since the battery is disconnected before the dielectric is inserted, the plate charge is unaffected by the dielectric and remains at 0.8 nC.

c. With the dielectric, the potential difference becomes

Equation:

$$V = \frac{1}{\kappa}V_0 = \frac{1}{2.1}40.0\,\text{V} = 19.0\,\text{V}.$$

d. The stored energy without the dielectric is

Equation:

$$U_0 = \frac{1}{2}C_0V_0^2 = \frac{1}{2}(20.0\,\text{pF})(40.0\,\text{V})^2 = 16.0\,\text{nJ}.$$

With the dielectric inserted, we use [link] to find that the stored energy decreases to

Equation:

$$U = \frac{1}{\kappa}U_0 = \frac{1}{2.1}16.0\,\text{nJ} = 7.6\,\text{nJ}.$$

Significance

Notice that the effect of a dielectric on the capacitance of a capacitor is a drastic increase of its capacitance. This effect is far more profound than a mere change in the geometry of a capacitor.

Note:
Exercise:

Problem:

Check Your Understanding When a dielectric is inserted into an isolated and charged capacitor, the stored energy decreases to 33% of its original value. (a) What is the dielectric constant? (b) How does the capacitance change?

Solution:

a. 3.0; b. $C = 3.0\,C_0$

Summary

- The capacitance of an empty capacitor is increased by a factor of κ when the space between its plates is completely filled by a dielectric with dielectric constant κ.
- Each dielectric material has its specific dielectric constant.
- The energy stored in an empty isolated capacitor is decreased by a factor of κ when the space between its plates is completely filled with a dielectric with dielectric constant κ.

Conceptual Questions

Exercise:

Problem:

Discuss what would happen if a conducting slab rather than a dielectric were inserted into the gap between the capacitor plates.

Solution:

answers may vary

Exercise:

Problem:

Discuss how the energy stored in an empty but charged capacitor changes when a dielectric is inserted if (a) the capacitor is isolated so that its charge does not change; (b) the capacitor remains connected to a battery so that the potential difference between its plates does not change.

Problems

Exercise:

Problem:

Show that for a given dielectric material, the maximum energy a parallel-plate capacitor can store is directly proportional to the volume of dielectric.

Exercise:

Problem:

An air-filled capacitor is made from two flat parallel plates 1.0 mm apart. The inside area of each plate is $8.0\,\text{cm}^2$. (a) What is the capacitance of this set of plates? (b) If the region between the plates is filled with a material whose dielectric constant is 6.0, what is the new capacitance?

Solution:

a. 7.1 pF; b. 42 pF

Exercise:

Problem:

A capacitor is made from two concentric spheres, one with radius 5.00 cm, the other with radius 8.00 cm. (a) What is the capacitance of this set of conductors? (b) If the region between the conductors is filled with a material whose dielectric constant is 6.00, what is the capacitance of the system?

Exercise:

Problem:

A parallel-plate capacitor has charge of magnitude $9.00\ \mu\text{C}$ on each plate and capacitance $3.00\ \mu\text{F}$ when there is air between the plates. The plates are separated by 2.00 mm. With the charge on the plates kept constant, a dielectric with $\kappa = 5$ is inserted between the plates, completely filling the volume between the plates. (a) What is the potential difference between the plates of the capacitor, before and after the dielectric has been inserted? (b) What is the electrical field at the point midway between the plates before and after the dielectric is inserted?

Solution:

a. before 3.00 V; after 0.600 V; b. before 1500 V/m; after 300 V/m

Exercise:

Problem:

Some cell walls in the human body have a layer of negative charge on the inside surface. Suppose that the surface charge densities are $\pm 0.50 \times 10^{-3} \mathrm{C/m^2}$, the cell wall is $5.0 \times 10^{-9} \mathrm{m}$ thick, and the cell wall material has a dielectric constant of $\kappa = 5.4$. (a) Find the magnitude of the electric field in the wall between two charge layers. (b) Find the potential difference between the inside and the outside of the cell. Which is at higher potential? (c) A typical cell in the human body has volume $10^{-16} \mathrm{m^3}$. Estimate the total electrical field energy stored in the wall of a cell of this size when assuming that the cell is spherical. (*Hint*: Calculate the volume of the cell wall.)

Exercise:

Problem:

A parallel-plate capacitor with only air between its plates is charged by connecting the capacitor to a battery. The capacitor is then disconnected from the battery, without any of the charge leaving the plates. (a) A voltmeter reads 45.0 V when placed across the capacitor. When a dielectric is inserted between the plates, completely filling the space, the voltmeter reads 11.5 V. What is the dielectric constant of the material? (b) What will the voltmeter read if the dielectric is now pulled away out so it fills only one-third of the space between the plates?

Solution:

a. 3.91; b. 22.8 V

Glossary

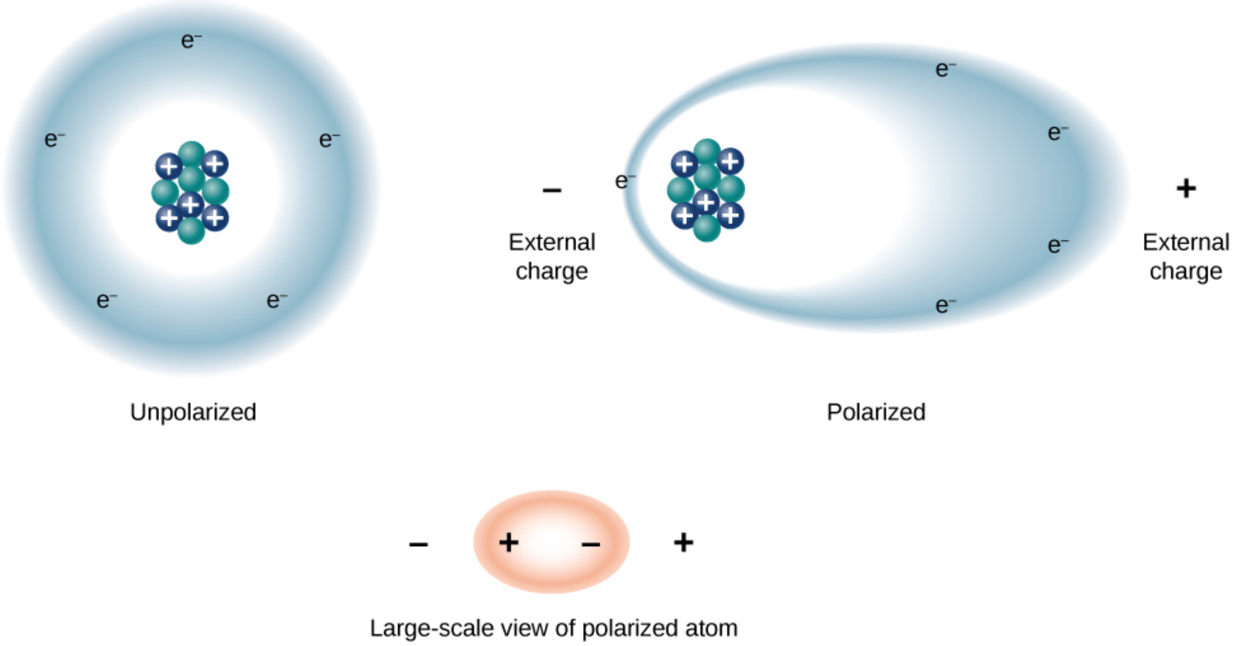
dielectric constant
: factor by which capacitance increases when a dielectric is inserted between the plates of a capacitor

Molecular Model of a Dielectric

By the end of this section, you will be able to:

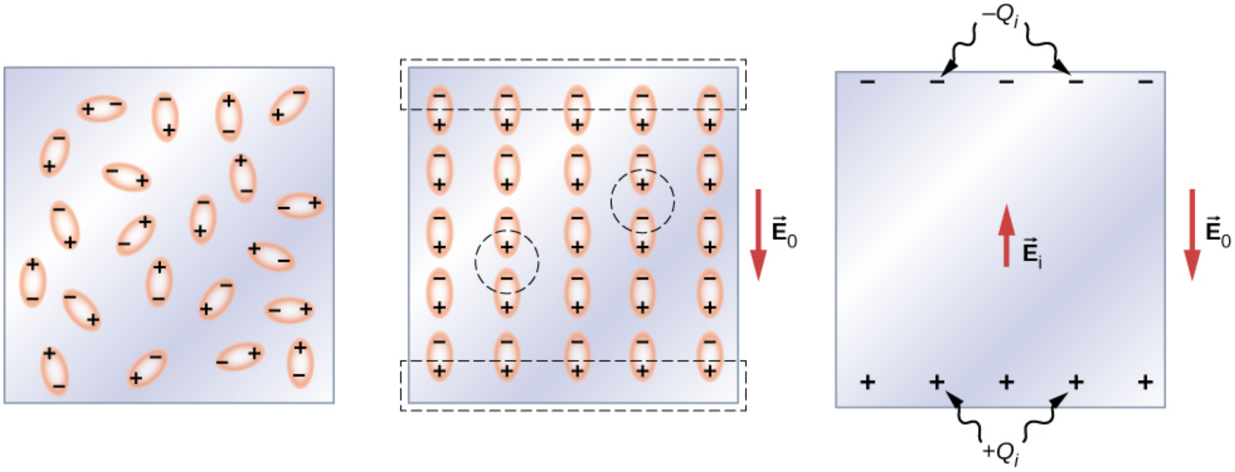
- Explain the polarization of a dielectric in a uniform electrical field
- Describe the effect of a polarized dielectric on the electrical field between capacitor plates
- Explain dielectric breakdown

We can understand the effect of a dielectric on capacitance by looking at its behavior at the molecular level. As we have seen in earlier chapters, in general, all molecules can be classified as either *polar* or *nonpolar*. There is a net separation of positive and negative charges in an isolated polar molecule, whereas there is no charge separation in an isolated nonpolar molecule ([link]). In other words, polar molecules have permanent *electric-dipole moments* and nonpolar molecules do not. For example, a molecule of water is polar, and a molecule of oxygen is nonpolar. Nonpolar molecules can become polar in the presence of an external electrical field, which is called *induced polarization*.



The concept of polarization: In an unpolarized atom or molecule, a negatively charged electron cloud is evenly distributed around positively charged centers, whereas a polarized atom or molecule has an excess of negative charge at one side so that the other side has an excess of positive charge. However, the entire system remains electrically neutral. The charge polarization may be caused by an external electrical field. Some molecules and atoms are permanently polarized (electric dipoles) even in the absence of an external electrical field (polar molecules and atoms).

Let’s first consider a dielectric composed of polar molecules. In the absence of any external electrical field, the electric dipoles are oriented randomly, as illustrated in [link](a). However, if the dielectric is placed in an external electrical field $\vec{\mathbf{E}}_0$, the polar molecules align with the external field, as shown in part (b) of the figure. Opposite charges on adjacent dipoles within the volume of dielectric neutralize each other, so there is no net charge within the dielectric (see the dashed circles in part (b)). However, this is not the case very close to the upper and lower surfaces that border the dielectric (the region enclosed by the dashed rectangles in part (b)), where the alignment does produce a net charge. Since the external electrical field merely aligns the dipoles, the dielectric as a whole is neutral, and the surface charges induced on its opposite faces are equal and opposite. These **induced surface charges** $+Q_i$ and $-Q_i$ produce an additional electrical field $\vec{\mathbf{E}}_\text{i}$ (an **induced electrical field**), which *opposes* the external field $\vec{\mathbf{E}}_0$, as illustrated in part (c).



(a) (b) (c)

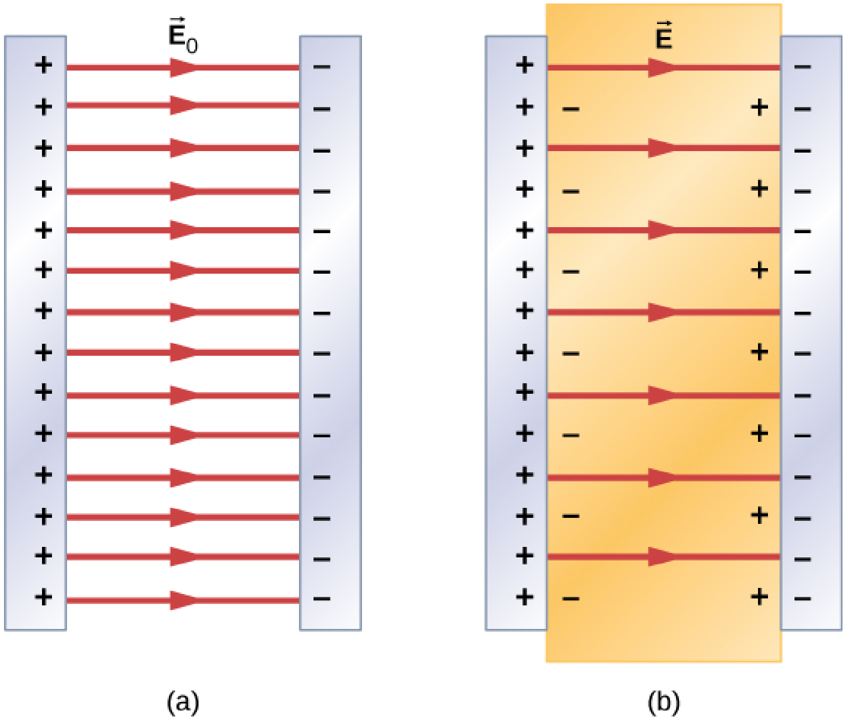
A dielectric with polar molecules: (a) In the absence of an external electrical field; (b) in the presence of an external electrical field $\vec{\mathbf{E}}_0$. The dashed lines indicate the regions immediately adjacent to the capacitor plates. (c) The induced electrical field $\vec{\mathbf{E}}_\text{i}$ inside the dielectric produced by the induced surface charge Q_i of the dielectric. Note that, in reality, the individual molecules are not perfectly aligned with an external field because of thermal fluctuations; however, the *average* alignment is along the field lines as shown.

The same effect is produced when the molecules of a dielectric are nonpolar. In this case, a nonpolar molecule acquires an **induced electric-dipole moment** because the external field $\vec{\mathbf{E}}_0$ causes a separation between its positive and negative charges. The induced dipoles of the nonpolar molecules align with $\vec{\mathbf{E}}_0$ in the same way as the permanent dipoles of the polar molecules are aligned (shown in part (b)). Hence, the electrical field within the dielectric is weakened regardless of whether its molecules are polar or nonpolar.

Therefore, when the region between the parallel plates of a charged capacitor, such as that shown in [link](a), is filled with a dielectric, within the dielectric there is an electrical field $\vec{\mathbf{E}}_0$ due to the *free charge* Q_0 on the capacitor plates and an electrical field $\vec{\mathbf{E}}_\text{i}$ due to the induced charge Q_i on the surfaces of the dielectric. Their vector sum gives the net electrical field $\vec{\mathbf{E}}$ within the dielectric between the capacitor plates (shown in part (b) of the figure):
Equation:

$$\vec{\mathbf{E}} = \vec{\mathbf{E}}_0 + \vec{\mathbf{E}}_\text{i}.$$

This net field can be considered to be the field produced by an *effective charge* $Q_0 - Q_i$ on the capacitor.



Electrical field: (a) In an empty capacitor, electrical field $\vec{\mathbf{E}}_0$. (b) In a dielectric-filled capacitor, electrical field $\vec{\mathbf{E}}$.

In most dielectrics, the net electrical field $\vec{\mathbf{E}}$ is proportional to the field $\vec{\mathbf{E}}_0$ produced by the free charge. In terms of these two electrical fields, the dielectric constant κ of the material is defined as

Note:
Equation:

$$\kappa = \frac{E_0}{E}.$$

Since $\vec{\mathbf{E}}_0$ and $\vec{\mathbf{E}}_\mathrm{i}$ point in opposite directions, the magnitude E is smaller than the magnitude E_0 and therefore $\kappa > 1$. Combining [link] with [link], and rearranging the terms, yields the following expression for the induced electrical field in a dielectric:

Note:
Equation:

$$\vec{\mathbf{E}}_i = \left(\frac{1}{\kappa} - 1\right)\vec{\mathbf{E}}_0.$$

When the magnitude of an external electrical field becomes too large, the molecules of dielectric material start to become ionized. A molecule or an atom is ionized when one or more electrons are removed from it and become free electrons, no longer bound to the molecular or atomic structure. When this happens, the material can conduct, thereby allowing charge to move through the dielectric from one capacitor plate to the other. This phenomenon is called **dielectric breakdown**. ([link] shows typical random-path patterns of electrical discharge during dielectric breakdown.) The critical value, E_c, of the electrical field at which the molecules of an insulator become ionized is called the **dielectric strength** of the material. The dielectric strength imposes a limit on the voltage that can be applied for a given plate separation in a capacitor. For example, the dielectric strength of air is $E_c = 3.0\,\text{MV/m}$, so for an air-filled capacitor with a plate separation of $d = 1.00\,\text{mm}$, the limit on the potential difference that can be safely applied across its plates without causing dielectric breakdown is $V = E_c d = (3.0 \times 10^6\,\text{V/m})(1.00 \times 10^{-3}\,\text{m}) = 3.0\,\text{kV}$.

However, this limit becomes 60.0 kV when the same capacitor is filled with Teflon™, whose dielectric strength is about $60.0\,\text{MV/m}$. Because of this limit imposed by the dielectric strength, the amount of charge that an air-filled capacitor can store is only $Q_0 = \kappa_{\text{air}} C_0 (3.0\,\text{kV})$ and the charge stored on the same Teflon™-filled capacitor can be as much as

Equation:

$$Q = \kappa_{\text{teflon}} C_0 (60.0\,\text{kV}) = \kappa_{\text{teflon}} \frac{Q_0}{\kappa_{\text{air}}(3.0\,\text{kV})}(60.0\,\text{kV}) = 20\frac{\kappa_{\text{teflon}}}{\kappa_{\text{air}}} Q_0 = 20\frac{2.1}{1.00059} Q_0 \cong 42\,Q_0,$$

which is about 42 times greater than a charge stored on an air-filled capacitor. Typical values of dielectric constants and dielectric strengths for various materials are given in [link]. Notice that the dielectric constant κ is exactly 1.0 for a vacuum (the empty space serves as a reference condition) and very close to 1.0 for air under normal conditions (normal pressure at room temperature). These two values are so close that, in fact, the properties of an air-filled capacitor are essentially the same as those of an empty capacitor.

Material	**Dielectric constant κ**	**Dielectric strength $E_c[\times 10^6 \text{V/m}]$**
Vacuum	1	∞
Dry air (1 atm)	1.00059	3.0
Teflon™	2.1	60 to 173
Paraffin	2.3	11

Material	**Dielectric constant κ**	**Dielectric strength $E_c[\times 10^6 \mathrm{V/m}]$**
Silicon oil	2.5	10 to 15
Polystyrene	2.56	19.7
Nylon	3.4	14
Paper	3.7	16
Fused quartz	3.78	8
Glass	4 to 6	9.8 to 13.8
Concrete	4.5	–
Bakelite	4.9	24
Diamond	5.5	2,000
Pyrex glass	5.6	14
Mica	6.0	118
Neoprene rubber	6.7	15.7 to 26.7
Water	80	—
Sulfuric acid	84 to 100	—
Titanium dioxide	86 to 173	–
Strontium titanate	310	8
Barium titanate	1,200 to 10,000	–
Calcium copper titanate	> 250,000	–

Representative Values of Dielectric Constants and Dielectric Strengths of Various Materials at Room Temperature

Not all substances listed in the table are good insulators, despite their high dielectric constants. Water, for example, consists of polar molecules and has a large dielectric constant of about 80. In a water molecule, electrons are more likely found around the oxygen nucleus than around the hydrogen nuclei. This makes the oxygen end of the molecule slightly negative and leaves the hydrogens end slightly positive, which makes the molecule easy to align along an external electrical field, and thus water has a large dielectric constant. However, the polar nature of water molecules also makes water a good solvent for many substances, which produces undesirable effects, because any concentration of free ions in water conducts electricity.

Example:

Electrical Field and Induced Surface Charge

Suppose that the distance between the plates of the capacitor in [link] is 2.0 mm and the area of each plate is $4.5 \times 10^{-3}\ \text{m}^2$. Determine: (a) the electrical field between the plates before and after the Teflon™ is inserted, and (b) the surface charge induced on the Teflon™ surfaces.

Strategy

In part (a), we know that the voltage across the empty capacitor is $V_0 = 40\ \text{V}$, so to find the electrical fields we use the relation $V = Ed$ and [link]. In part (b), knowing the magnitude of the electrical field, we use the expression for the magnitude of electrical field near a charged plate $E = \sigma/\varepsilon_0$, where σ is a uniform surface charge density caused by the surface charge. We use the value of free charge $Q_0 = 8.0 \times 10^{-10}\ \text{C}$ obtained in [link].

Solution

a. The electrical field E_0 between the plates of an empty capacitor is
Equation:

$$E_0 = \frac{V_0}{d} = \frac{40\ \text{V}}{2.0 \times 10^{-3}\text{m}} = 2.0 \times 10^4\ \text{V/m}.$$

The electrical field E with the Teflon™ in place is
Equation:

$$E = \frac{1}{\kappa}E_0 = \frac{1}{2.1}2.0 \times 10^4\ \text{V/m} = 9.5 \times 10^3\ \text{V/m}.$$

b. The effective charge on the capacitor is the difference between the free charge Q_0 and the induced charge Q_i. The electrical field in the Teflon™ is caused by this effective charge. Thus
Equation:

$$E = \frac{1}{\varepsilon_0}\sigma = \frac{1}{\varepsilon_0}\frac{Q_0 - Q_\text{i}}{A}.$$

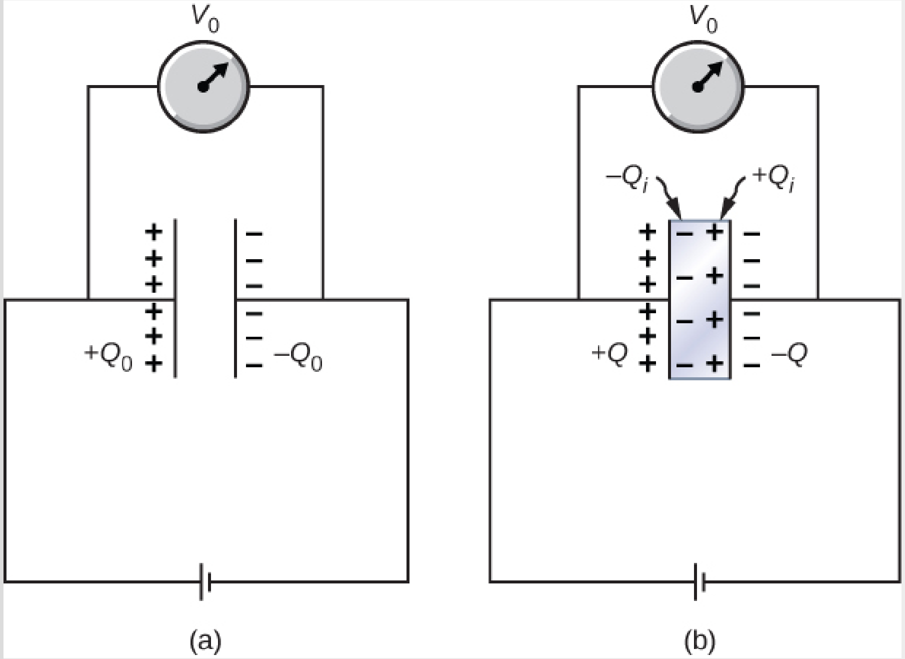
We invert this equation to obtain Q_i, which yields
Equation:

$$\begin{aligned} Q_\text{i} &= Q_0 - \varepsilon_0 AE \\ &= 8.0 \times 10^{-10}\text{C} - \left(8.85 \times 10^{-12}\ \frac{\text{C}^2}{\text{N}\cdot\text{m}^2}\right)\left(4.5 \times 10^{-3}\ \text{m}^2\right)\left(9.5 \times 10^3\ \frac{\text{V}}{\text{m}}\right) \\ &= 4.2 \times 10^{-10}\text{C} = 0.42\ \text{nC}. \end{aligned}$$

Example:

Inserting a Dielectric into a Capacitor Connected to a Battery

When a battery of voltage V_0 is connected across an empty capacitor of capacitance C_0, the charge on its plates is Q_0, and the electrical field between its plates is E_0. A dielectric of dielectric constant κ is inserted between the plates *while the battery remains in place*, as shown in [link]. (a) Find the capacitance C, the voltage V across the capacitor, and the electrical field E between the plates after the dielectric is inserted. (b) Obtain an expression for the free charge Q on the plates of the filled capacitor and the induced charge Q_{i} on the dielectric surface in terms of the original plate charge Q_0.



A dielectric is inserted into the charged capacitor while the capacitor remains connected to the battery.

Strategy

We identify the known values: V_0, C_0,E_0, κ, and Q_0. Our task is to express the unknown values in terms of these known values.

Solution

(a) The capacitance of the filled capacitor is $C = \kappa C_0$. Since the battery is always connected to the capacitor plates, the potential difference between them does not change; hence, $V = V_0$. Because of that, the electrical field in the filled capacitor is the same as the field in the empty capacitor, so we can obtain directly that

Equation:

$$E = \frac{V}{d} = \frac{V_0}{d} = E_0.$$

(b) For the filled capacitor, the free charge on the plates is

Equation:

$$Q = CV = (\kappa C_0)V_0 = \kappa(C_0V_0) = \kappa Q_0.$$

The electrical field E in the filled capacitor is due to the effective charge $Q - Q_{\text{i}}$ ([link](b)). Since $E = E_0$, we have

Equation:

$$\frac{Q - Q_{\text{i}}}{\varepsilon_0 A} = \frac{Q_0}{\varepsilon_0 A}.$$

Solving this equation for Q_{i}, we obtain for the induced charge

Equation:

$$Q_{\text{i}} = Q - Q_0 = \kappa Q_0 - Q_0 = (\kappa - 1)Q_0.$$

Significance

Notice that for materials with dielectric constants larger than 2 (see [link]), the induced charge on the surface of dielectric is larger than the charge on the plates of a vacuum capacitor. The opposite is true for gasses like air whose dielectric constant is smaller than 2.

Note:

Exercise:

Problem:

Check Your Understanding Continuing with [link], show that when the battery is connected across the plates the energy stored in dielectric-filled capacitor is $U = \kappa U_0$ (larger than the energy U_0 of an empty capacitor kept at the same voltage). Compare this result with the result $U = U_0/\kappa$ found previously for an isolated, charged capacitor.

Note:

Exercise:

Problem:

Check Your Understanding Repeat the calculations of [link] for the case in which the battery remains connected while the dielectric is placed in the capacitor.

Solution:

a. $C_0 = 20\,\text{pF}$, $C = 42\,\text{pF}$; b. $Q_0 = 0.8\,\text{nC}$, $Q = 1.7\,\text{nC}$; c. $V_0 = V = 40\,\text{V}$; d. $U_0 = 16\,\text{nJ}$, $U = 34\,\text{nJ}$

Summary

- When a dielectric is inserted between the plates of a capacitor, equal and opposite surface charge is induced on the two faces of the dielectric. The induced surface charge produces an induced electrical field that opposes the field of the free charge on the capacitor plates.
- The dielectric constant of a material is the ratio of the electrical field in vacuum to the net electrical field in the material. A capacitor filled with dielectric has a larger capacitance than an empty capacitor.
- The dielectric strength of an insulator represents a critical value of electrical field at which the molecules in an insulating material start to become ionized. When this happens, the material can conduct and dielectric breakdown is observed.

Key Equations

Capacitance	$C = \frac{Q}{V}$
Capacitance of a parallel-plate capacitor	$C = \varepsilon_0 \frac{A}{d}$
Capacitance of a vacuum spherical capacitor	$C = 4\pi\varepsilon_0 \frac{R_1 R_2}{R_2 - R_1}$
Capacitance of a vacuum cylindrical capacitor	$C = \frac{2\pi\varepsilon_0 l}{\ln(R_2/R_1)}$
Capacitance of a series combination	$\frac{1}{C_\text{S}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots$
Capacitance of a parallel combination	$C_\text{P} = C_1 + C_2 + C_3 + \cdots$
Energy density	$u_E = \frac{1}{2}\varepsilon_0 E^2$
Energy stored in a capacitor	$U_C = \frac{1}{2}V^2 C = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}QV$
Capacitance of a capacitor with dielectric	$C = \kappa C_0$
Energy stored in an isolated capacitor with dielectric	$U = \frac{1}{\kappa}U_0$
Dielectric constant	$\kappa = \frac{E_0}{E}$
Induced electrical field in a dielectric	$\vec{\mathbf{E}}_\text{i} = \left(\frac{1}{\kappa} - 1\right)\vec{\mathbf{E}}_0$

Conceptual Questions

Exercise:

Problem: Distinguish between dielectric strength and dielectric constant.

Solution:

Dielectric strength is a critical value of an electrical field above which an insulator starts to conduct; a dielectric constant is the ratio of the electrical field in vacuum to the net electrical field in a material.

Exercise:

Problem: Water is a good solvent because it has a high dielectric constant. Explain.

Exercise:

Problem:

Water has a high dielectric constant. Explain why it is then not used as a dielectric material in capacitors.

Solution:

Water is a good solvent.

Exercise:

Problem:

Elaborate on why molecules in a dielectric material experience net forces on them in a non-uniform electrical field but not in a uniform field.

Exercise:

Problem:

Explain why the dielectric constant of a substance containing permanent molecular electric dipoles decreases with increasing temperature.

Solution:

When energy of thermal motion is large (high temperature), an electrical field must be large too in order to keep electric dipoles aligned with it.

Exercise:

Problem:

Give a reason why a dielectric material increases capacitance compared with what it would be with air between the plates of a capacitor. How does a dielectric material also allow a greater voltage to be applied to a capacitor? (The dielectric thus increases C and permits a greater V.)

Exercise:

Problem:

Elaborate on the way in which the polar character of water molecules helps to explain water’s relatively large dielectric constant.

Solution:

answers may vary

Exercise:

Problem:

Sparks will occur between the plates of an air-filled capacitor at a lower voltage when the air is humid than when it is dry. Discuss why, considering the polar character of water molecules.

Problems

Exercise:

Problem:

Two flat plates containing equal and opposite charges are separated by material 4.0 mm thick with a dielectric constant of 5.0. If the electrical field in the dielectric is 1.5 MV/m, what are (a) the charge density on the capacitor plates, and (b) the induced charge density on the surfaces of the dielectric?

Exercise:

Problem:

For a Teflon™-filled, parallel-plate capacitor, the area of the plate is $50.0\;\mathrm{cm}^2$ and the spacing between the plates is 0.50 mm. If the capacitor is connected to a 200-V battery, find (a) the free charge on the capacitor plates, (b) the electrical field in the dielectric, and (c) the induced charge on the dielectric surfaces.

Solution:

a. 37 nC; b. 0.4 MV/m; c. 19 nC

Exercise:

Problem:

Find the capacitance of a parallel-plate capacitor having plates with a surface area of $5.00\,m^2$ and separated by 0.100 mm of Teflon™.

Exercise:

Problem:

(a) What is the capacitance of a parallel-plate capacitor with plates of area $1.50\;\mathrm{m}^2$ that are separated by 0.0200 mm of neoprene rubber? (b) What charge does it hold when 9.00 V is applied to it?

Solution:

a. 4.4 µF; b. $4.0\;\times\;10^{-5}$ C

Exercise:

Problem:

Two parallel plates have equal and opposite charges. When the space between the plates is evacuated, the electrical field is $E=3.20\;\times\;10^5\;\mathrm{V/m}$. When the space is filled with dielectric, the electrical field is $E=2.50\;\times\;10^5\;\mathrm{V/m}$. (a) What is the surface charge density on each surface of the dielectric? (b) What is the dielectric constant?

Exercise:

Problem:

The dielectric to be used in a parallel-plate capacitor has a dielectric constant of 3.60 and a dielectric strength of $1.60\;\times\;10^7\;\mathrm{V/m}$. The capacitor has to have a capacitance of 1.25 nF and must be able to withstand a maximum potential difference 5.5 kV. What is the minimum area the plates of the capacitor may have?

Solution:

$0.0135\ \mathrm{m}^2$

Exercise:

Problem:

When a 360-nF air capacitor is connected to a power supply, the energy stored in the capacitor is $18.5\ \mu\mathrm{J}$. While the capacitor is connected to the power supply, a slab of dielectric is inserted that completely fills the space between the plates. This increases the stored energy by $23.2\ \mu\mathrm{J}$. (a) What is the potential difference between the capacitor plates? (b) What is the dielectric constant of the slab?

Exercise:

Problem:

A parallel-plate capacitor has square plates that are 8.00 cm on each side and 3.80 mm apart. The space between the plates is completely filled with two square slabs of dielectric, each 8.00 cm on a side and 1.90 mm thick. One slab is Pyrex glass and the other slab is polystyrene. If the potential difference between the plates is 86.0 V, find how much electrical energy can be stored in this capacitor.

Solution:

$0.185\ \mu\mathrm{J}$

Additional Problems

Exercise:

Problem:

A capacitor is made from two flat parallel plates placed 0.40 mm apart. When a charge of $0.020\ \mathrm{\mu C}$ is placed on the plates the potential difference between them is 250 V. (a) What is the capacitance of the plates? (b) What is the area of each plate? (c) What is the charge on the plates when the potential difference between them is 500 V? (d) What maximum potential difference can be applied between the plates so that the magnitude of electrical fields between the plates does not exceed 3.0 MV/m?

Exercise:

Problem:

An air-filled (empty) parallel-plate capacitor is made from two square plates that are 25 cm on each side and 1.0 mm apart. The capacitor is connected to a 50-V battery and fully charged. It is then disconnected from the battery and its plates are pulled apart to a separation of 2.00 mm. (a) What is the capacitance of this new capacitor? (b) What is the charge on each plate? (c) What is the electrical field between the plates?

Solution:

a. 0.277 nF; b. 27.7 nC; c. 50 kV/m

Exercise:

Problem:

Suppose that the capacitance of a variable capacitor can be manually changed from 100 to 800 pF by turning a dial connected to one set of plates by a shaft, from 0° to 180°. With the dial set at 180° (corresponding to $C = 800\,\text{pF}$), the capacitor is connected to a 500-V source. After charging, the capacitor is disconnected from the source, and the dial is turned to 0°. (a) What is the charge on the capacitor? (b) What is the voltage across the capacitor when the dial is set to 0°?

Exercise:

Problem:

Earth can be considered as a spherical capacitor with two plates, where the negative plate is the surface of Earth and the positive plate is the bottom of the ionosphere, which is located at an altitude of approximately 70 km. The potential difference between Earth's surface and the ionosphere is about 350,000 V. (a) Calculate the capacitance of this system. (b) Find the total charge on this capacitor. (c) Find the energy stored in this system.

Solution:

a. 0.065 F; b. 23,000 C; c. 4.0 GJ

Exercise:

Problem:

A $4.00\text{-}\mu\text{F}$ capacitor and a $6.00\text{-}\mu\text{F}$ capacitor are connected in parallel across a 600-V supply line. (a) Find the charge on each capacitor and voltage across each. (b) The charged capacitors are disconnected from the line and from each other. They are then reconnected to each other with terminals of unlike sign together. Find the final charge on each capacitor and the voltage across each.

Exercise:

Problem:

Three capacitors having capacitances of 8.40, 8.40, and 4.20 μF, respectively, are connected in series across a 36.0-V potential difference. (a) What is the charge on the $4.20\text{-}\mu\text{F}$ capacitor? (b) The capacitors are disconnected from the potential difference without allowing them to discharge. They are then reconnected in parallel with each other with the positively charged plates connected together. What is the voltage across each capacitor in the parallel combination?

Solution:

a. $75.6\,\mu\text{C}$; b. 10.8 V

Exercise:

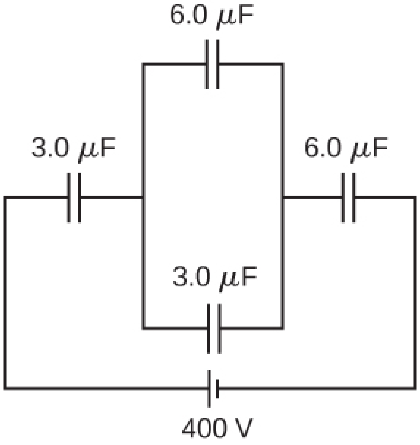
Problem:

A parallel-plate capacitor with capacitance $5.0\,\mu\text{F}$ is charged with a 12.0-V battery, after which the battery is disconnected. Determine the minimum work required to increase the separation between the plates by a factor of 3.

Exercise:

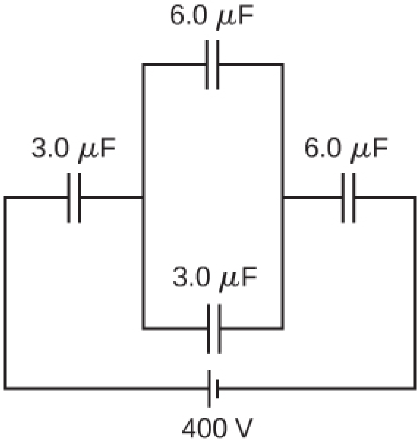
Problem:

(a) How much energy is stored in the electrical fields in the capacitors (in total) shown below? (b) Is this energy equal to the work done by the 400-V source in charging the capacitors?



Solution:

a. 0.13 J; b. no, because of resistive heating in connecting wires that is always present, but the circuit schematic does not indicate resistors



Exercise:

Problem:

Three capacitors having capacitances 8.4, 8.4, and 4.2 μF are connected in series across a 36.0-V potential difference. (a) What is the total energy stored in all three capacitors? (b) The capacitors are disconnected from the potential difference without allowing them to discharge. They are then reconnected in parallel with each other with the positively charged plates connected together. What is the total energy now stored in the capacitors?

Exercise:

Problem:

(a) An $8.00\text{-}\mu\mathrm{F}$ capacitor is connected in parallel to another capacitor, producing a total capacitance of $5.00\ \mu\mathrm{F}$. What is the capacitance of the second capacitor? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. $-3.00\ \mathrm{\mu F}$; b. You cannot have a negative C_2 capacitance. c. The assumption that they were hooked up in parallel, rather than in series, is incorrect. A parallel connection always produces a greater capacitance, while here a smaller capacitance was assumed. This could only happen if the capacitors are connected in series.

Exercise:

Problem:

(a) On a particular day, it takes $9.60\ \times\ 10^3\ \mathrm{J}$ of electrical energy to start a truck's engine. Calculate the capacitance of a capacitor that could store that amount of energy at 12.0 V. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem:

(a) A certain parallel-plate capacitor has plates of area $4.00\ \mathrm{m}^2$, separated by 0.0100 mm of nylon, and stores 0.170 C of charge. What is the applied voltage? (b) What is unreasonable about this result? (c) Which assumptions are responsible or inconsistent?

Solution:

a. 14.2 kV; b. The voltage is unreasonably large, more than 100 times the breakdown voltage of nylon. c. The assumed charge is unreasonably large and cannot be stored in a capacitor of these dimensions.

Exercise:

Problem:

A prankster applies 450 V to an $80.0\text{-}\mu\mathrm{F}$ capacitor and then tosses it to an unsuspecting victim. The victim's finger is burned by the discharge of the capacitor through 0.200 g of flesh. Estimate, what is the temperature increase of the flesh? Is it reasonable to assume that no thermodynamic phase change happened?

Challenge Problems

Exercise:

Problem:

A spherical capacitor is formed from two concentric spherical conducting spheres separated by vacuum. The inner sphere has radius 12.5 cm and the outer sphere has radius 14.8 cm. A potential difference of 120 V is applied to the capacitor. (a) What is the capacitance of the capacitor? (b) What is the magnitude of the electrical field at $r = 12.6\ \text{cm}$, just outside the inner sphere? (c) What is the magnitude of the electrical field at $r = 14.7\ \text{cm}$, just inside the outer sphere? (d) For a parallel-plate capacitor the electrical field is uniform in the region between the plates, except near the edges of the plates. Is this also true for a spherical capacitor?

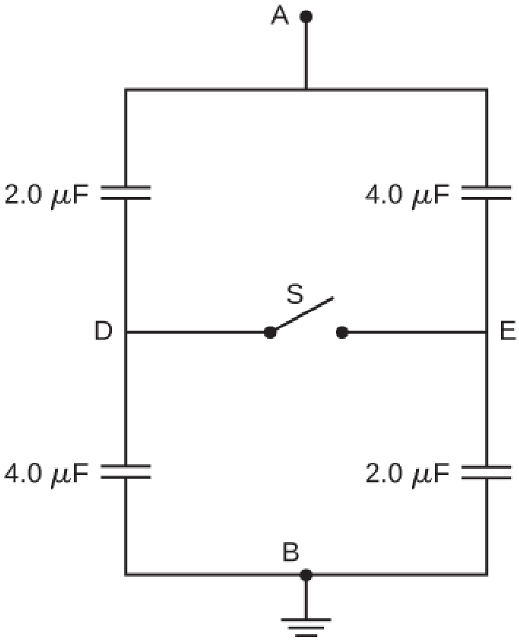
Solution:

a. 89.6 pF; b. 6.09 kV/m; c. 4.47 kV/m; d. no

Exercise:

Problem:

The network of capacitors shown below are all uncharged when a 300-V potential is applied between points A and B with the switch S open. (a) What is the potential difference $V_E - V_D$? (b) What is the potential at point E after the switch is closed? (c) How much charge flows through the switch after it is closed?



Exercise:

Problem:

Electronic flash units for cameras contain a capacitor for storing the energy used to produce the flash. In one such unit the flash lasts for 1/675 fraction of a second with an average light power output of 270 kW. (a) If the conversion of electrical energy to light is 95% efficient (because the rest of the energy goes to thermal energy), how much energy must be stored in the capacitor for one flash? (b) The capacitor has a potential difference between its plates of 125 V when the stored energy equals the value stored in part (a). What is the capacitance?

Solution:

a. 421 J; b. 53.9 mF

Exercise:

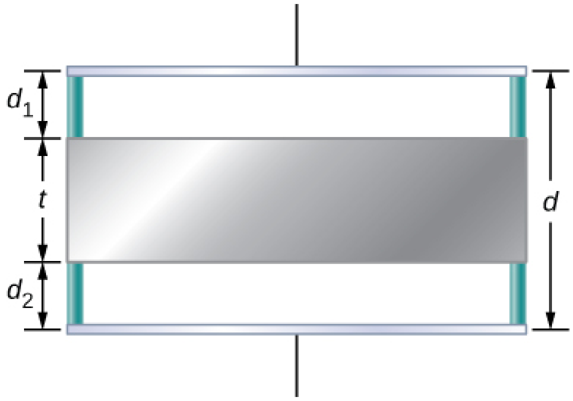
Problem:

A spherical capacitor is formed from two concentric spherical conducting shells separated by a vacuum. The inner sphere has radius 12.5 cm and the outer sphere has radius 14.8 cm. A potential difference of 120 V is applied to the capacitor. (a) What is the energy density at $r = 12.6\ \text{cm}$, just outside the inner sphere? (b) What is the energy density at $r = 14.7\ \text{cm}$, just inside the outer sphere? (c) For the parallel-plate capacitor the energy density is uniform in the region between the plates, except near the edges of the plates. Is this also true for the spherical capacitor?

Exercise:

Problem:

A metal plate of thickness *t* is held in place between two capacitor plates by plastic pegs, as shown below. The effect of the pegs on the capacitance is negligible. The area of each capacitor plate and the area of the top and bottom surfaces of the inserted plate are all *A*. What is the capacitance of this system?



Solution:

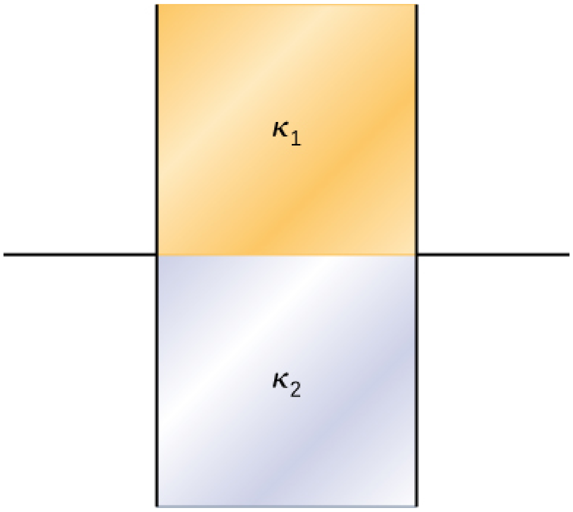
$C = \varepsilon_0 A/(d_1 + d_2)$

Exercise:

Problem:

A parallel-plate capacitor is filled with two dielectrics, as shown below. When the plate area is *A* and separation between plates is *d*, show that the capacitance is given by

$$C = \varepsilon_0 \frac{A}{d} \frac{\kappa_1 + \kappa_2}{2}.$$

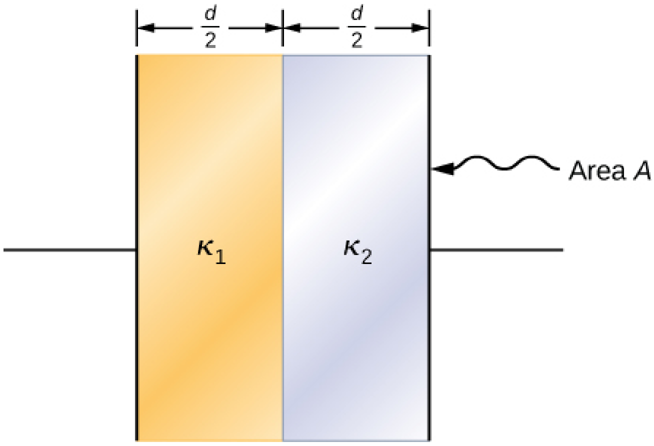


Exercise:

Problem:

A parallel-plate capacitor is filled with two dielectrics, as shown below. Show that the capacitance is given by

$C=2\varepsilon_0\frac{A}{d}\,\frac{\kappa_1\kappa_2}{\kappa_1+\kappa_2}$.



Solution:

proof

Exercise:

Problem:

A capacitor has parallel plates of area $12\ \text{cm}^2$ separated by 2.0 mm. The space between the plates is filled with polystyrene. (a) Find the maximum permissible voltage across the capacitor to avoid dielectric breakdown. (b) When the voltage equals the value found in part (a), find the surface charge density on the surface of the dielectric.

Glossary

dielectric breakdown
phenomenon that occurs when an insulator becomes a conductor in a strong electrical field

dielectric strength
critical electrical field strength above which molecules in insulator begin to break down and the insulator starts to conduct

induced electric-dipole moment
dipole moment that a nonpolar molecule may acquire when it is placed in an electrical field

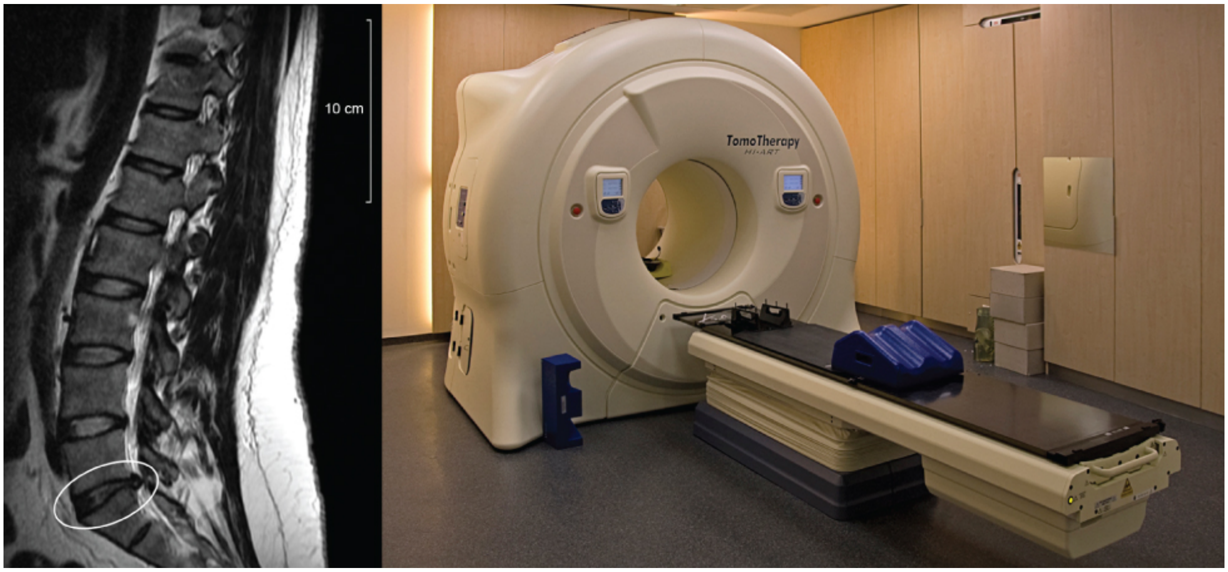
induced electrical field
electrical field in the dielectric due to the presence of induced charges

induced surface charges
charges that occur on a dielectric surface due to its polarization

Introduction class="introduction"

Magnetic resonance imaging (MRI) uses superconducting magnets and produces high-resolution images without the danger of radiation. The image on the left shows the spacing of vertebrae along a human spinal column, with the circle indicating where the vertebrae are too close due to a ruptured disc. On the right is a picture of the MRI instrument, which surrounds the patient on all sides. A large amount of electrical current is

required to
operate the
electromagnets
(credit right:
modification of
work by “digital
cat”/Flickr).



In this chapter, we study the electrical current through a material, where the electrical current is the rate of flow of charge. We also examine a characteristic of materials known as the resistance. Resistance is a measure of how much a material impedes the flow of charge, and it will be shown that the resistance depends on temperature. In general, a good conductor, such as copper, gold, or silver, has very low resistance. Some materials, called superconductors, have zero resistance at very low temperatures.

High currents are required for the operation of electromagnets. Superconductors can be used to make electromagnets that are 10 times stronger than the strongest conventional electromagnets. These superconducting magnets are used in the construction of magnetic resonance imaging (MRI) devices that can be used to make high-resolution images of the human body. The chapter-opening picture shows an MRI image of the vertebrae of a human subject and the MRI device itself.

Superconducting magnets have many other uses. For example, superconducting magnets are used in the Large Hadron Collider (LHC) to curve the path of protons in the ring.

Electrical Current
By the end of this section, you will be able to:

- Describe an electrical current
- Define the unit of electrical current
- Explain the direction of current flow

Up to now, we have considered primarily static charges. When charges did move, they were accelerated in response to an electrical field created by a voltage difference. The charges lost potential energy and gained kinetic energy as they traveled through a potential difference where the electrical field did work on the charge.

Although charges do not require a material to flow through, the majority of this chapter deals with understanding the movement of charges through a material. The rate at which the charges flow past a location—that is, the amount of charge per unit time—is known as the *electrical current*. When charges flow through a medium, the current depends on the voltage applied, the material through which the charges flow, and the state of the material. Of particular interest is the motion of charges in a conducting wire. In previous chapters, charges were accelerated due to the force provided by an electrical field, losing potential energy and gaining kinetic energy. In this chapter, we discuss the situation of the force provided by an electrical field in a conductor, where charges lose kinetic energy to the material reaching a constant velocity, known as the “*drift velocity*.” This is analogous to an object falling through the atmosphere and losing kinetic energy to the air, reaching a constant terminal velocity.

If you have ever taken a course in first aid or safety, you may have heard that in the event of electric shock, it is the current, not the voltage, which is the important factor on the severity of the shock and the amount of damage to the human body. Current is measured in units called amperes; you may have noticed that circuit breakers in your home and fuses in your car are rated in amps (or amperes). But what is the ampere and what does it measure?

Defining Current and the Ampere

Electrical current is defined to be the rate at which charge flows. When there is a large current present, such as that used to run a refrigerator, a large amount of charge moves through the wire in a small amount of time. If the current is small, such as that used to operate a handheld calculator, a small amount of charge moves through the circuit over a long period of time.

Note:
Electrical Current
The average electrical current I is the rate at which charge flows,
Equation:

$$I_{\text{ave}} = \frac{\Delta Q}{\Delta t},$$

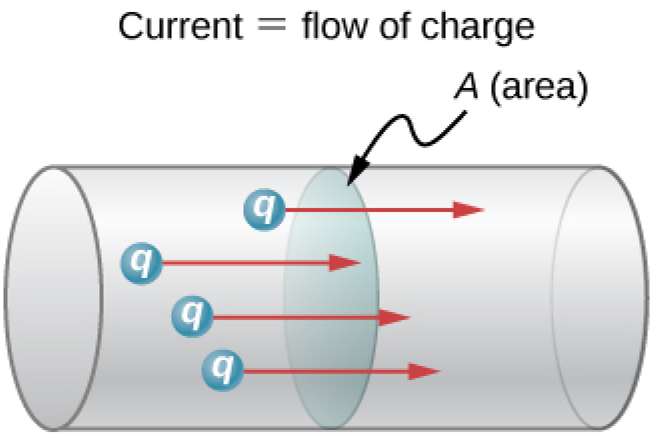
where ΔQ is the amount of net charge passing through a given cross-sectional area in time Δt ([link]). The SI unit for current is the **ampere** (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \frac{\Delta Q}{\Delta t}$, we see that an ampere is defined as one coulomb of charge passing through a given area per second:
Equation:

$$1\text{A} \equiv 1\frac{\text{C}}{\text{s}}.$$

The instantaneous electrical current, or simply the **electrical current**, is the time derivative of the charge that flows and is found by taking the limit of the average electrical current as $\Delta t \to 0$:
Equation:

$$I = \lim_{\Delta t \to 0} \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt}.$$

Most electrical appliances are rated in amperes (or amps) required for proper operation, as are fuses and circuit breakers.



The rate of flow of charge is current. An ampere is the flow of one coulomb of charge through an area in one second. A current of one amp would result from 6.25×10^{18} electrons flowing through the area *A* each second.

Example:

Calculating the Average Current

The main purpose of a battery in a car or truck is to run the electric starter motor, which starts the engine. The operation of starting the vehicle requires a large current to be supplied by the battery. Once the engine starts, a device called an alternator takes over supplying the electric power required for running the vehicle and for charging the battery.

(a) What is the average current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine? (b) How long does it take 1.00 C of charge to flow from the battery?

Strategy

We can use the definition of the average current in the equation $I = \frac{\Delta Q}{\Delta t}$ to find the average current in part (a), since charge and time are given. For part (b), once we know the average current, we can its definition $I = \frac{\Delta Q}{\Delta t}$ to find the time required for 1.00 C of charge to flow from the battery.

Solution

a. Entering the given values for charge and time into the definition of current gives

Equation:

$$I = \frac{\Delta Q}{\Delta t} = \frac{720\ \text{C}}{4.00\ \text{s}} = 180\ \text{C/s} = 180\ \text{A}.$$

b. Solving the relationship $I = \frac{\Delta Q}{\Delta t}$ for time Δt and entering the known values for charge and current gives

Equation:

$$\Delta t = \frac{\Delta Q}{I} = \frac{1.00\ \text{C}}{180\ \text{C/s}} = 5.56\ \times\ 10^{-3}\text{s} = 5.56\ \text{ms}.$$

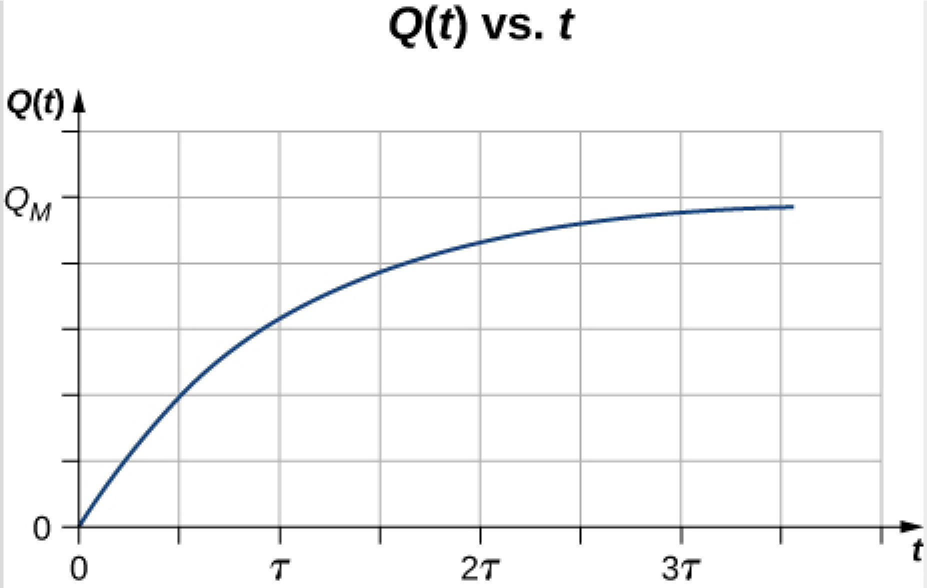
Significance

a. This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these “starter motors” are fairly large to overcome the inertia of the engine. b. A high current requires a short time to supply a large amount of charge. This large current is needed to supply the large amount of energy needed to start the engine.

Example:

Calculating Instantaneous Currents

Consider a charge moving through a cross-section of a wire where the charge is modeled as $Q\left(t\right) = Q_M\left(1 - e^{-t/\tau}\right)$. Here, Q_M is the charge after a long period of time, as time approaches infinity, with units of coulombs, and τ is a time constant with units of seconds (see [link]). What is the current through the wire?



A graph of the charge moving through a cross-section of a wire over time.

Strategy

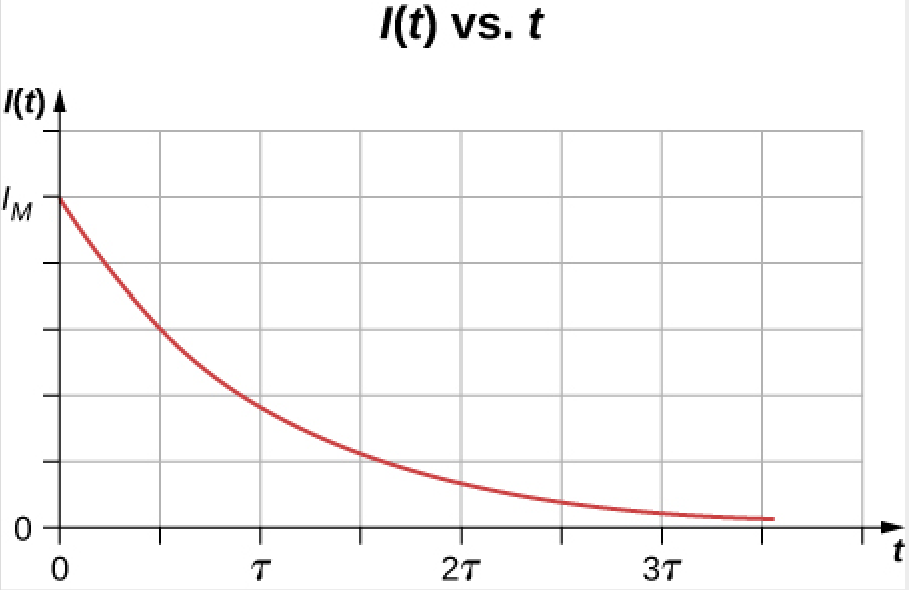
The current through the cross-section can be found from $I=\frac{dQ}{dt}$. Notice from the figure that the charge increases to Q_M and the derivative decreases, approaching zero, as time increases ([link]).

Solution

The derivative can be found using $\frac{d}{dx}e^u=e^u\frac{du}{dx}$.

Equation:

$$I=\frac{dQ}{dt}=\frac{d}{dt}\left[Q_M\left(1-e^{-t/\tau}\right)\right]=\frac{Q_M}{\tau}e^{-t/\tau}.$$



A graph of the current flowing through the wire over time.

Significance
The current through the wire in question decreases exponentially, as shown in [link]. In later chapters, it will be shown that a time-dependent current appears when a capacitor charges or discharges through a resistor. Recall that a capacitor is a device that stores charge. You will learn about the resistor in Model of Conduction in Metals.

Note:
Exercise:

Problem:

Check Your Understanding Handheld calculators often use small solar cells to supply the energy required to complete the calculations needed to complete your next physics exam. The current needed to run your calculator can be as small as 0.30 mA. How long would it take for 1.00 C of charge to flow from the solar cells? Can solar cells be used, instead of batteries, to start traditional internal combustion engines presently used in most cars and trucks?

Solution:

The time for 1.00 C of charge to flow would be $\Delta t = \frac{\Delta Q}{I} = \frac{1.00\ \text{C}}{0.300 \times 10^{-3}\ \text{C/s}} = 3.33 \times 10^3\text{s}$, slightly less than an hour. This is quite different from the 5.55 ms for the truck battery. The calculator takes a very small amount of energy to operate, unlike the truck’s starter motor. There are several reasons that vehicles use batteries and not solar cells. Aside from the obvious fact that a light source to run the solar cells for a car or truck is not always available, the large amount of current needed to start the engine cannot easily be supplied by present-day solar cells. Solar cells can possibly be used to charge the batteries. Charging the battery requires a small amount of energy when compared to the energy required to run the engine and the other accessories such as the heater and air conditioner. Present day solar-powered cars are powered by solar panels, which may power an electric motor, instead of an internal combustion engine.

Note:
Exercise:

Problem:

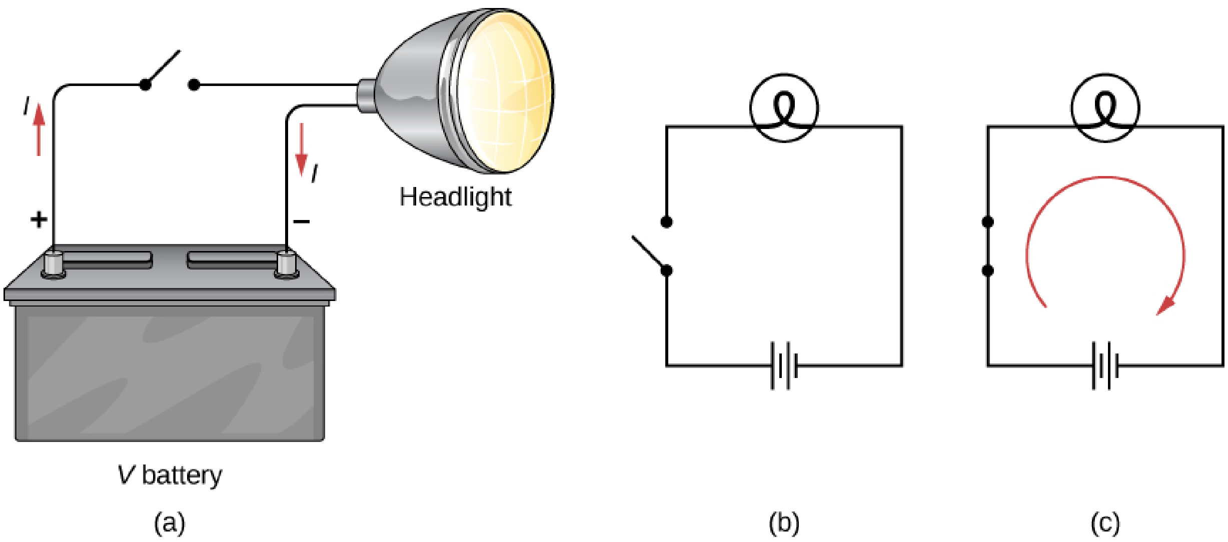
Check Your Understanding Circuit breakers in a home are rated in amperes, normally in a range from 10 amps to 30 amps, and are used to protect the residents from harm and their appliances from damage due to large currents. A single 15-amp circuit breaker may be used to protect several outlets in the living room, whereas a single 20-amp circuit breaker may be used to protect the refrigerator in the kitchen. What can you deduce from this about current used by the various appliances?

Solution:

The total current needed by all the appliances in the living room (a few lamps, a television, and your laptop) draw less current and require less power than the refrigerator.

Current in a Circuit

In the previous paragraphs, we defined the current as the charge that flows through a cross-sectional area per unit time. In order for charge to flow through an appliance, such as the headlight shown in [link], there must be a complete path (or **circuit**) from the positive terminal to the negative terminal. Consider a simple circuit of a car battery, a switch, a headlight lamp, and wires that provide a current path between the components. In order for the lamp to light, there must be a complete path for current flow. In other words, a charge must be able to leave the positive terminal of the battery, travel through the component, and back to the negative terminal of the battery. The switch is there to control the circuit. Part (a) of the figure shows the simple circuit of a car battery, a switch, a conducting path, and a headlight lamp. Also shown is the **schematic** of the circuit [part (b)]. A schematic is a graphical representation of a circuit and is very useful in visualizing the main features of a circuit. Schematics use standardized symbols to represent the components in a circuits and solid lines to represent the wires connecting the components. The battery is shown as a series of long and short lines, representing the historic voltaic pile. The lamp is shown as a circle with a loop inside, representing the filament of an incandescent bulb. The switch is shown as two points with a conducting bar to connect the two points and the wires connecting the components are shown as solid lines. The schematic in part (c) shows the direction of current flow when the switch is closed.

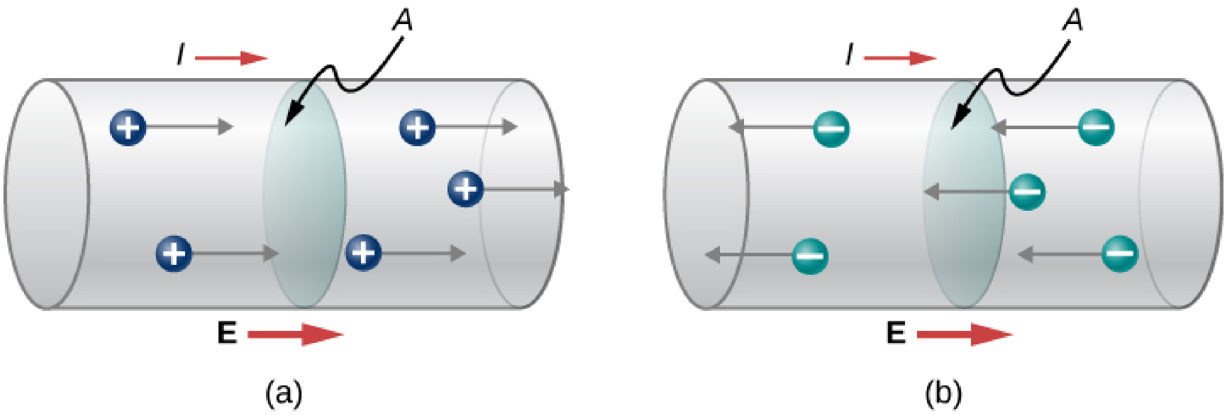


(a) A simple electric circuit of a headlight (lamp), a battery, and a switch. When the switch is closed, an uninterrupted path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by parallel lines, which resemble plates in the original design of a battery. The longer lines indicate the positive terminal. The conducting wires are shown as solid lines. The switch is shown, in the open position, as two terminals with a line representing a conducting bar that can make contact between the two terminals. The lamp is represented by a circle encompassing a filament, as would be seen in an incandescent light bulb. (c) When the switch is closed, the circuit is complete and current flows from the positive terminal to the negative terminal of the battery.

When the switch is closed in [link](c), there is a complete path for charges to flow, from the positive terminal of the battery, through the switch, then through the headlight and back to the negative terminal of the battery. Note that the direction of current flow is from positive to negative. The direction of **conventional current** is always represented in the direction that positive charge would flow, from the positive terminal to the negative terminal.

The conventional current flows from the positive terminal to the negative terminal, but depending on the actual situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de Graaff generator, used for nuclear research, can produce a current of pure positive charges, such as protons. In the Tevatron Accelerator at Fermilab, before it was shut down in 2011, beams of protons and antiprotons traveling in opposite directions were collided. The protons are positive and therefore their current is in the same direction as they travel. The antiprotons are negativity charged and thus their current is in the opposite direction that the actual particles travel.

A closer look at the current flowing through a wire is shown in [link]. The figure illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American scientist and statesman Benjamin Franklin in the 1700s. Having no knowledge of the particles that make up the atom (namely the proton, electron, and neutron), Franklin believed that electrical current flowed from a material that had more of an “electrical fluid” and to a material that had less of this “electrical fluid.” He coined the term *positive* for the material that had more of this electrical fluid and *negative* for the material that lacked the electrical fluid. He surmised that current would flow from the material with more electrical fluid—the positive material—to the negative material, which has less electrical fluid. Franklin called this direction of current a positive current flow. This was pretty advanced thinking for a man who knew nothing about the atom.



Current *I* is the rate at which charge moves through an area *A*, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electrical field. (a) Positive charges move in the direction of the electrical field, which is the same direction as conventional current. (b) Negative charges move in the direction opposite to the electrical field. Conventional current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

We now know that a material is positive if it has a greater number of protons than electrons, and it is negative if it has a greater number of electrons than protons. In a conducting metal, the current flow is due primarily to electrons flowing from the negative material to the positive material, but for historical reasons, we consider the positive current flow and the current is shown to flow from the positive terminal of the battery to the negative terminal.

It is important to realize that an electrical field is present in conductors and is responsible for producing the current ([link]). In previous chapters, we considered the static electrical case, where charges in a conductor quickly redistribute themselves on the surface of the conductor in order to cancel out the external electrical field and restore equilibrium. In the case of an electrical circuit, the charges are prevented from ever reaching equilibrium by an external source of electric potential, such as a battery. The energy needed to move the charge is supplied by the electric potential from the battery.

Although the electrical field is responsible for the motion of the charges in the conductor, the work done on the charges by the electrical field does not increase the kinetic energy of the charges. We will show that the electrical field is responsible for keeping the electric charges moving at a “drift velocity.”

Summary

- The average electrical current I_{ave} is the rate at which charge flows, given by $I_{\text{ave}} = \frac{\Delta Q}{\Delta t}$, where ΔQ is the amount of charge passing through an area in time Δt.
- The instantaneous electrical current, or simply the current *I*, is the rate at which charge flows. Taking the limit as the change in time approaches zero, we have $I = \frac{dQ}{dt}$, where $\frac{dQ}{dt}$ is the time derivative of the charge.
- The direction of conventional current is taken as the direction in which positive charge moves. In a simple direct-current (DC) circuit, this will be from the positive terminal of the battery to the negative terminal.
- The SI unit for current is the ampere, or simply the amp (A), where $1\text{ A} = 1\text{ C/s}$.
- Current consists of the flow of free charges, such as electrons, protons, and ions.

Conceptual Questions

Exercise:

Problem:

Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.

Solution:

If a wire is carrying a current, charges enter the wire from the voltage source’s positive terminal and leave at the negative terminal, so the total charge remains zero while the current flows through it.

Exercise:

Problem:

Car batteries are rated in ampere-hours $(\mathrm{A \cdot h})$. To what physical quantity do ampere-hours correspond (voltage, current, charge, energy, power,…)?

Exercise:

Problem:

When working with high-power electric circuits, it is advised that whenever possible, you work “one-handed” or “keep one hand in your pocket.” Why is this a sensible suggestion?

Solution:

Using one hand will reduce the possibility of “completing the circuit” and having current run through your body, especially current running through your heart.

Problems

Exercise:

Problem:

A Van de Graaff generator is one of the original particle accelerators and can be used to accelerate charged particles like protons or electrons. You may have seen it used to make human hair stand on end or produce large sparks. One application of the Van de Graaff generator is to create X-rays by bombarding a hard metal target with the beam. Consider a beam of protons at 1.00 keV and a current of 5.00 mA produced by the generator. (a) What is the speed of the protons? (b) How many protons are produced each second?

Solution:

a. $v = 4.38 \times 10^5 \frac{\text{m}}{\text{s}}$;
b. $\Delta q = 5.00 \times 10^{-3}\text{C}, \quad \text{no. of protons} = 3.13 \times 10^{16}$

Exercise:

Problem:

A cathode ray tube (CRT) is a device that produces a focused beam of electrons in a vacuum. The electrons strike a phosphor-coated glass screen at the end of the tube, which produces a bright spot of light. The position of the bright spot of light on the screen can be adjusted by deflecting the electrons with electrical fields, magnetic fields, or both. Although the CRT tube was once commonly found in televisions, computer displays, and oscilloscopes, newer appliances use a liquid crystal display (LCD) or plasma screen. You still may come across a CRT in your study of science. Consider a CRT with an electron beam average current of $25.00\mu\text{ A}$. How many electrons strike the screen every minute?

Exercise:

Problem:

How many electrons flow through a point in a wire in 3.00 s if there is a constant current of $I = 4.00\text{ A}$?

Solution:

$I = \frac{\Delta Q}{\Delta t}, \quad \Delta Q = 12.00\text{ C}$
no. of electrons = 7.5×10^{19}

Exercise:

Problem:

A conductor carries a current that is decreasing exponentially with time. The current is modeled as $I = I_0 e^{-t/\tau}$, where $I_0 = 3.00\text{ A}$ is the current at time $t = 0.00\text{ s}$ and $\tau = 0.50\text{ s}$ is the time constant. How much charge flows through the conductor between $t = 0.00\text{ s}$ and $t = 3\tau$?

Exercise:

Problem:

The quantity of charge through a conductor is modeled as $Q = 4.00\frac{\text{C}}{\text{s}^4}t^4 - 1.00\frac{\text{C}}{\text{s}}t + 6.00\text{ mC}$.

What is the current at time $t = 3.00\text{ s}$?

Solution:

$$I(t) = 0.016\frac{\text{C}}{\text{s}^4}t^3 - 0.001\frac{\text{C}}{\text{s}}$$
$$I(3.00\text{ s}) = 0.431\text{ A}$$

Exercise:

Problem:

The current through a conductor is modeled as $I(t) = I_m \sin\left(2\pi\left[60\text{ Hz}\right]t\right)$. Write an equation for the charge as a function of time.

Exercise:

Problem:

The charge on a capacitor in a circuit is modeled as $Q(t) = Q_{\max}\cos\left(\omega t + \phi\right)$. What is the current through the circuit as a function of time?

Solution:

$$I(t) = -I_{\max} \sin\,(\omega t + \phi)$$

Glossary

ampere (amp)
SI unit for current; $1\text{ A} = 1\text{ C/s}$

circuit
complete path that an electrical current travels along

conventional current
current that flows through a circuit from the positive terminal of a battery through the circuit to the negative terminal of the battery

electrical current
rate at which charge flows, $I = \frac{dQ}{dt}$

schematic
graphical representation of a circuit using standardized symbols for components and solid lines for the wire connecting the components

Model of Conduction in Metals

By the end of this section, you will be able to:

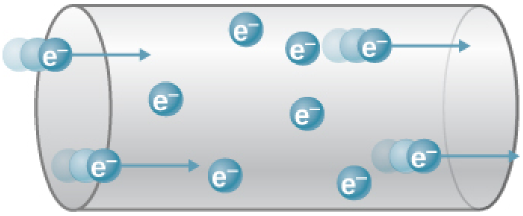
- Define the drift velocity of charges moving through a metal
- Define the vector current density
- Describe the operation of an incandescent lamp

When electrons move through a conducting wire, they do not move at a constant velocity, that is, the electrons do not move in a straight line at a constant speed. Rather, they interact with and collide with atoms and other free electrons in the conductor. Thus, the electrons move in a zig-zag fashion and drift through the wire. We should also note that even though it is convenient to discuss the direction of current, current is a scalar quantity. When discussing the velocity of charges in a current, it is more appropriate to discuss the current density. We will come back to this idea at the end of this section.

Drift Velocity

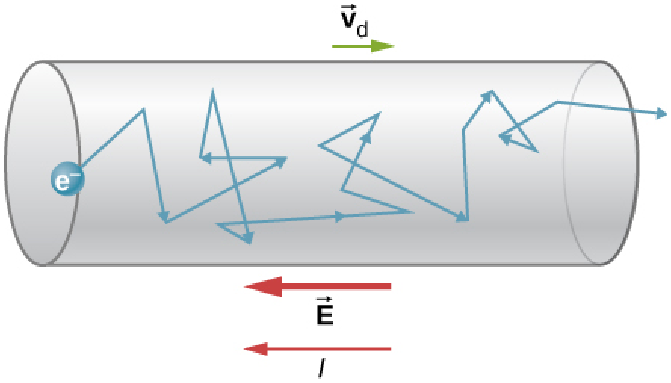
Electrical signals move very rapidly. Telephone conversations carried by currents in wires cover large distances without noticeable delays. Lights come on as soon as a light switch is moved to the 'on' position. Most electrical signals carried by currents travel at speeds on the order of $10^8\ \text{m/s}$, a significant fraction of the speed of light. Interestingly, the individual charges that make up the current move much slower on average, typically drifting at speeds on the order of 10^{-4}m/s. How do we reconcile these two speeds, and what does it tell us about standard conductors?

The high speed of electrical signals results from the fact that the force between charges acts rapidly at a distance. Thus, when a free charge is forced into a wire, as in [link], the incoming charge pushes other charges ahead of it due to the repulsive force between like charges. These moving charges push on charges farther down the line. The density of charge in a system cannot easily be increased, so the signal is passed on rapidly. The resulting electrical shock wave moves through the system at nearly the speed of light. To be precise, this fast-moving signal, or shock wave, is a rapidly propagating change in the electrical field.



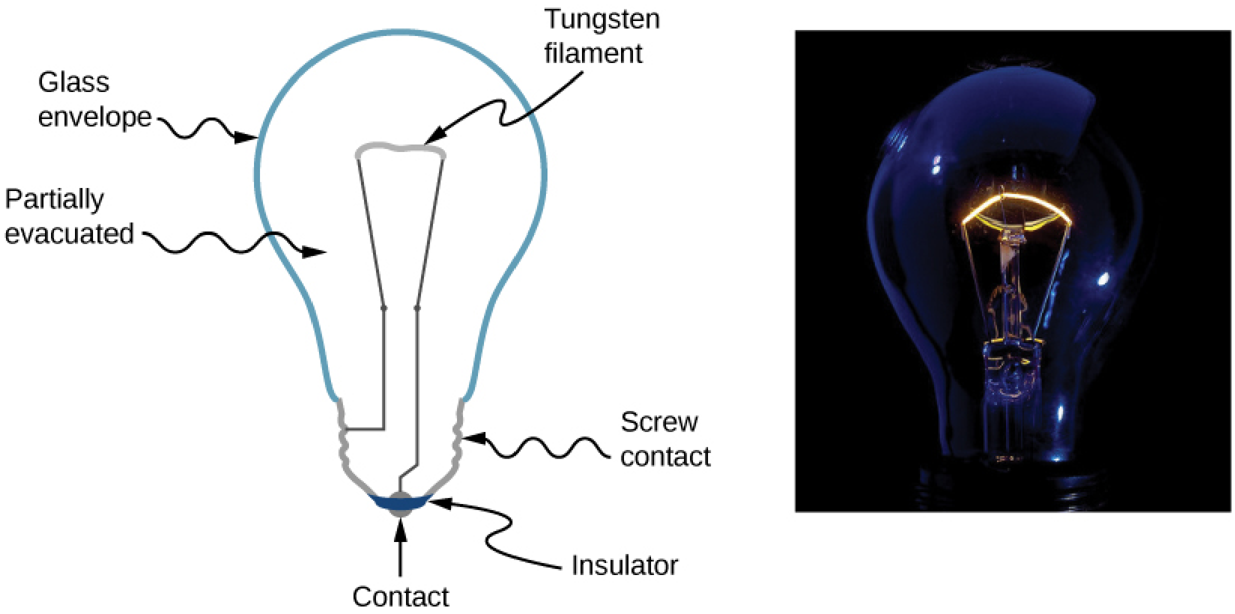
When charged particles are forced into this volume of a conductor, an equal number are quickly forced to leave. The repulsion between like charges makes it difficult to increase the number of charges in a volume. Thus, as one charge enters, another leaves almost immediately, carrying the signal rapidly forward.

Good conductors have large numbers of free charges. In metals, the free charges are free electrons. (In fact, good electrical conductors are often good heat conductors too, because large numbers of free electrons can transport thermal energy as well as carry electrical current.) [link] shows how free electrons move through an ordinary conductor. The distance that an individual electron can move between collisions with atoms or other electrons is quite small. The electron paths thus appear nearly random, like the motion of atoms in a gas. But there is an electrical field in the conductor that causes the electrons to drift in the direction shown (opposite to the field, since they are negative). The **drift velocity** $\vec{\mathbf{v}}_\mathrm{d}$ is the average velocity of the free charges. Drift velocity is quite small, since there are so many free charges. If we have an estimate of the density of free electrons in a conductor, we can calculate the drift velocity for a given current. The larger the density, the lower the velocity required for a given current.



Free electrons moving in a conductor make many collisions with other electrons and other particles. A typical path of one electron is shown. The average velocity of the free charges is called the drift velocity $\vec{\mathbf{v}}_\mathrm{d}$ and for electrons, it is in the direction opposite to the electrical field. The collisions normally transfer energy to the conductor, requiring a constant supply of energy to maintain a steady current.

Free-electron collisions transfer energy to the atoms of the conductor. The electrical field does work in moving the electrons through a distance, but that work does not increase the kinetic energy (nor speed) of the electrons. The work is transferred to the conductor’s atoms, often increasing temperature. Thus, a continuous power input is required to keep a current flowing. (An exception is superconductors, for reasons we shall explore in a later chapter. Superconductors can have a steady current without a continual supply of energy—a great energy savings.) For a conductor that is not a superconductor, the supply of energy can be useful, as in an incandescent light bulb filament ([link]). The supply of energy is necessary to increase the temperature of the tungsten filament, so that the filament glows.



The incandescent lamp is a simple design. A tungsten filament is placed in a partially evacuated glass envelope. One end of the filament is attached to the screw base, which is made out of a conducting material. The second end of the filament is attached to a second contact in the base of the bulb. The two contacts are separated by an insulating material. Current flows through the filament, and the temperature of the filament becomes large enough to cause the filament to glow and produce light. However, these bulbs are not very energy efficient, as evident from the heat coming from the bulb. In the year 2012, the United States, along with many other countries, began to phase out incandescent lamps in favor of more energy-efficient lamps, such as light-emitting diode (LED) lamps and compact fluorescent lamps (CFL) (credit right: modification of work by Serge Saint).

We can obtain an expression for the relationship between current and drift velocity by considering the number of free charges in a segment of wire, as illustrated in [link]. The number of free charges per unit volume, or the number density of free charges, is given the symbol n where $n=\frac{\text{number of charges}}{\text{volume}}$. The value of n depends on the material. The shaded segment has a volume $Av_{\text{d}}dt$, so that the number of free charges in the volume is $nAv_{\text{d}}dt$. The charge dQ in this segment is thus $qnAv_{\text{d}}dt$, where q is the amount of charge on each carrier. (The magnitude of the charge of electrons is $q=1.60\ \times\ 10^{-19}$ C.) Current is charge moved per unit time; thus, if all the original charges move out of this segment in time dt, the current is

Equation:

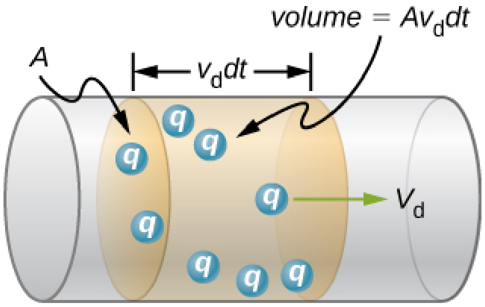
$$I=\frac{dQ}{dt}=qnAv_{\text{d}}.$$

Rearranging terms gives

Note:
Equation:

$$v_{\text{d}} = \frac{I}{nqA}$$

where v_{d} is the drift velocity, *n* is the free charge density, *A* is the cross-sectional area of the wire, and *I* is the current through the wire. The carriers of the current each have charge *q* and move with a drift velocity of magnitude v_{d}.

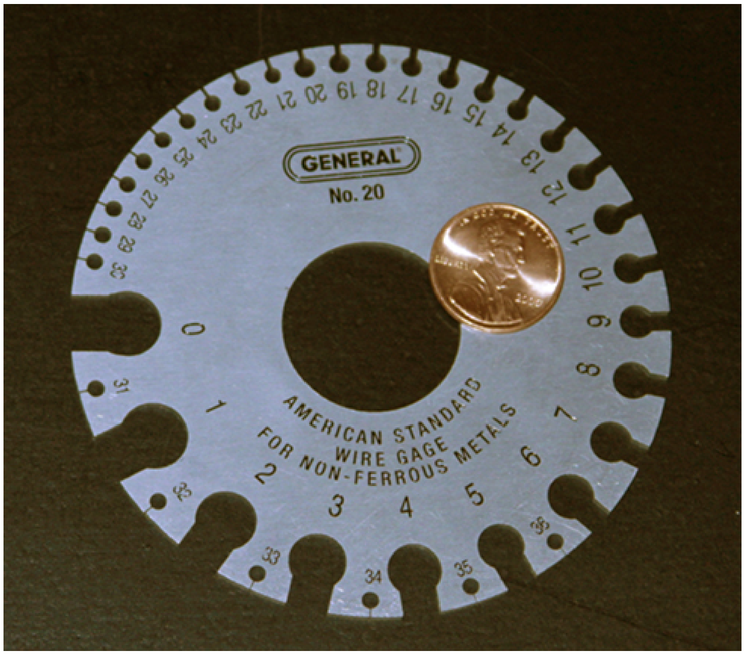


All the charges in the shaded volume of this wire move out in a time *dt*, having a drift velocity of magnitude v_{d}.

Note that simple drift velocity is not the entire story. The speed of an electron is sometimes much greater than its drift velocity. In addition, not all of the electrons in a conductor can move freely, and those that do move might move somewhat faster or slower than the drift velocity. So what do we mean by free electrons?

Atoms in a metallic conductor are packed in the form of a lattice structure. Some electrons are far enough away from the atomic nuclei that they do not experience the attraction of the nuclei as strongly as the inner electrons do. These are the free electrons. They are not bound to a single atom but can instead move freely among the atoms in a “sea” of electrons. When an electrical field is applied, these free electrons respond by accelerating. As they move, they collide with the atoms in the lattice and with other electrons, generating thermal energy, and the conductor gets warmer. In an insulator, the organization of the atoms and the structure do not allow for such free electrons.

As you know, electric power is usually supplied to equipment and appliances through round wires made of a conducting material (copper, aluminum, silver, or gold) that are stranded or solid. The diameter of the wire determines the current-carrying capacity—the larger the diameter, the greater the current-carrying capacity. Even though the current-carrying capacity is determined by the diameter, wire is not normally characterized by the diameter directly. Instead, wire is commonly sold in a unit known as “gauge.” Wires are manufactured by passing the material through circular forms called “drawing dies.” In order to make thinner wires, manufacturers draw the wires through multiple dies of successively thinner diameter. Historically, the gauge of the wire was related to the number of drawing processes required to manufacture the wire. For this reason, the larger the gauge, the smaller the diameter. In the United States, the American Wire Gauge (AWG) was developed to standardize the system. Household wiring commonly consists of 10-gauge (2.588-mm diameter) to 14-gauge (1.628-mm diameter) wire. A device used to measure the gauge of wire is shown in [link].



A device for measuring the gauge of electrical wire. As you can see, higher gauge numbers indicate thinner wires. (credit: Joseph J. Trout)

Example:

Calculating Drift Velocity in a Common Wire

Calculate the drift velocity of electrons in a copper wire with a diameter of 2.053 mm (12-gauge) carrying a 20.0-A current, given that there is one free electron per copper atom. (Household wiring often contains 12-gauge copper wire, and the maximum current allowed in such wire is usually 20.0 A.) The density of copper is $8.80 \times 10^3 \text{ kg/m}^3$ and the atomic mass of copper is 63.54 g/mol.

Strategy

We can calculate the drift velocity using the equation $I = nqAv_{\mathrm{d}}$. The current is $I = 20.00\ \mathrm{A}$ and $q = 1.60\ \times\ 10^{-19}\mathrm{C}$ is the charge of an electron. We can calculate the area of a cross-section of the wire using the formula $A = \pi r^2$, where *r* is one-half the diameter. The given diameter is 2.053 mm, so *r* is 1.0265 mm. We are given the density of copper, $8.80\ \times\ 10^3\ \mathrm{kg/m}^3$, and the atomic mass of copper is $63.54\ \mathrm{g/mol}$. We can use these two quantities along with Avogadro's number, $6.02\ \times\ 10^{23}\ \mathrm{atoms/mol}$, to determine *n*, the number of free electrons per cubic meter.

Solution

First, we calculate the density of free electrons in copper. There is one free electron per copper atom. Therefore, the number of free electrons is the same as the number of copper atoms per m^3. We can now find *n* as follows:

Equation:

$$\begin{aligned} n &= \frac{1\,e^-}{\text{atom}} \times \frac{6.02 \times 10^{23}\text{atoms}}{\text{mol}} \times \frac{1\ \text{mol}}{63.54\ \text{g}} \times \frac{1000\ \text{g}}{\text{kg}} \times \frac{8.80 \times 10^3\text{kg}}{1\ \text{m}^3} \\ &= 8.34\ \times\ 10^{28} e^-/\mathrm{m}^3. \end{aligned}$$

The cross-sectional area of the wire is

Equation:

$$A = \pi r^2 = \pi \left(\frac{2.05\ \times\ 10^{-3}\mathrm{m}}{2}\right)^2 = 3.30\ \times\ 10^{-6}\mathrm{m}^2.$$

Rearranging $I = nqAv_{\mathrm{d}}$ to isolate drift velocity gives

Equation:

$$v_{\mathrm{d}} = \frac{I}{nqA} = \frac{20.00\ \mathrm{A}}{(8.34\ \times\ 10^{28}/\mathrm{m}^3)(-1.60\ \times\ 10^{-19}\mathrm{C})(3.30\ \times\ 10^{-6}\mathrm{m}^2)} = -4.54\ \times\ 10^{-4}\ \mathrm{m/s}.$$

Significance

The minus sign indicates that the negative charges are moving in the direction opposite to conventional current. The small value for drift velocity (on the order of $10^{-4}\ \mathrm{m/s}$) confirms that the signal moves on the order of 10^{12} times faster (about $10^8\ \mathrm{m/s}$) than the charges that carry it.

Note:

Exercise:

Problem:

Check Your Understanding In [link], the drift velocity was calculated for a 2.053-mm diameter (12-gauge) copper wire carrying a 20-amp current. Would the drift velocity change for a 1.628-mm diameter (14-gauge) wire carrying the same 20-amp current?

Solution:

The diameter of the 14-gauge wire is smaller than the diameter of the 12-gauge wire. Since the drift velocity is inversely proportional to the cross-sectional area, the drift velocity in the 14-gauge wire is larger than the drift velocity in the 12-gauge wire carrying the same current. The number of electrons per cubic meter will remain constant.

Current Density

Although it is often convenient to attach a negative or positive sign to indicate the overall direction of motion of the charges, current is a scalar quantity, $I = \frac{dQ}{dt}$. It is often necessary to discuss the details of the motion of the charge, instead of discussing the overall motion of the charges. In such cases, it is necessary to discuss the current density, $\vec{\mathbf{J}}$, a vector quantity. The **current density** is the flow of charge through an infinitesimal area, divided by the area. The current density must take into account the local magnitude and direction of the charge flow, which varies from point to point. The unit of current density is ampere per meter squared, and the direction is defined as the direction of net flow of positive charges through the area.

The relationship between the current and the current density can be seen in [link]. The differential current flow through the area $d\vec{\mathbf{A}}$ is found as

Equation:

$$dI = \vec{\mathbf{J}} \cdot d\vec{\mathbf{A}} = JdA\cos\theta,$$

where θ is the angle between the area and the current density. The total current passing through area $d\vec{\mathbf{A}}$ can be found by integrating over the area,

Note:

Equation:

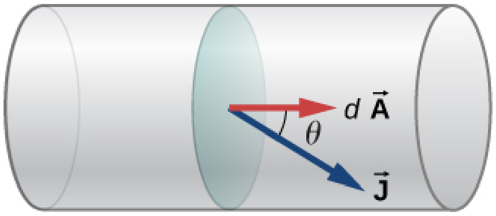
$$I = \iint_{\text{area}} \vec{\mathbf{J}} \cdot d\vec{\mathbf{A}}.$$

Consider the magnitude of the current density, which is the current divided by the area:

Equation:

$$J = \frac{I}{A} = \frac{n\,|q|Av_{\text{d}}}{A} = n\,|q|v_{\text{d}}.$$

Thus, the current density is $\vec{\mathbf{J}} = nq\vec{\mathbf{v}}_{\text{d}}$. If q is positive, $\vec{\mathbf{v}}_{\text{d}}$ is in the same direction as the electrical field $\vec{\mathbf{E}}$. If q is negative, $\vec{\mathbf{v}}_{\text{d}}$ is in the opposite direction of $\vec{\mathbf{E}}$. Either way, the direction of the current density $\vec{\mathbf{J}}$ is in the direction of the electrical field $\vec{\mathbf{E}}$.



The current density $\vec{\mathbf{J}}$ is defined as the current passing through an infinitesimal cross-sectional area divided by the area. The direction of the current density is the direction of the net flow of positive charges and the magnitude is equal to the current divided by the infinitesimal area.

Example:

Calculating the Current Density in a Wire

The current supplied to a lamp with a 100-W light bulb is 0.87 amps. The lamp is wired using a copper wire with diameter 2.588 mm (10-gauge). Find the magnitude of the current density.

Strategy

The current density is the current moving through an infinitesimal cross-sectional area divided by the area. We can calculate the magnitude of the current density using $J = \frac{I}{A}$. The current is given as 0.87 A. The cross-sectional area can be calculated to be $A = 5.26\,\text{mm}^2$.

Solution

Calculate the current density using the given current $I = 0.87\,\text{A}$ and the area, found to be $A = 5.26\,\text{mm}^2$.

Equation:

$$J = \frac{I}{A} = \frac{0.87\,\text{A}}{5.26 \times 10^{-6}\text{m}^2} = 1.65 \times 10^5 \frac{\text{A}}{\text{m}^2}.$$

Significance

The current density in a conducting wire depends on the current through the conducting wire and the cross-sectional area of the wire. For a given current, as the diameter of the wire increases, the charge density decreases.

Note:

Exercise:

Problem:

Check Your Understanding The current density is proportional to the current and inversely proportional to the area. If the current density in a conducting wire increases, what would happen to the drift velocity of the charges in the wire?

Solution:

The current density in a conducting wire increases due to an increase in current. The drift velocity is inversely proportional to the current $\left(v_d = \frac{I}{nqA}\right)$, so the drift velocity would decrease.

What is the significance of the current density? The current density is proportional to the current, and the current is the number of charges that pass through a cross-sectional area per second. The charges move through the conductor, accelerated by the electric force provided by the electrical field. The electrical field is created when a voltage is applied across the conductor. In Ohm's Law, we will use this relationship between the current density and the electrical field to examine the relationship between the current through a conductor and the voltage applied.

Summary

- The current through a conductor depends mainly on the motion of free electrons.
- When an electrical field is applied to a conductor, the free electrons in a conductor do not move through a conductor at a constant speed and direction; instead, the motion is almost random due to collisions with atoms and other free electrons.
- Even though the electrons move in a nearly random fashion, when an electrical field is applied to the conductor, the overall velocity of the electrons can be defined in terms of a drift velocity.
- The current density is a vector quantity defined as the current through an infinitesimal area divided by the area.
- The current can be found from the current density, $I = \iint_{\text{area}} \vec{\mathbf{J}} \cdot d\vec{\mathbf{A}}$.
- An incandescent light bulb is a filament of wire enclosed in a glass bulb that is partially evacuated. Current runs through the filament, where the electrical energy is converted to light and heat.

Conceptual Questions

Exercise:

Problem:

Incandescent light bulbs are being replaced with more efficient LED and CFL light bulbs. Is there any obvious evidence that incandescent light bulbs might not be that energy efficient? Is energy converted into anything but visible light?

Exercise:

Problem:

It was stated that the motion of an electron appears nearly random when an electrical field is applied to the conductor. What makes the motion nearly random and differentiates it from the random motion of molecules in a gas?

Solution:

Even though the electrons collide with atoms and other electrons in the wire, they travel from the negative terminal to the positive terminal, so they drift in one direction. Gas molecules travel in completely random directions.

Exercise:

Problem:

Electric circuits are sometimes explained using a conceptual model of water flowing through a pipe. In this conceptual model, the voltage source is represented as a pump that pumps water through pipes and the pipes connect components in the circuit. Is a conceptual model of water flowing through a pipe an adequate representation of the circuit? How are electrons and wires similar to water molecules and pipes? How are they different?

Exercise:

Problem: An incandescent light bulb is partially evacuated. Why do you suppose that is?

Solution:

In the early years of light bulbs, the bulbs are partially evacuated to reduce the amount of heat conducted through the air to the glass envelope. Dissipating the heat would cool the filament, increasing the amount of energy needed to produce light from the filament. It also protects the glass from the heat produced from the hot filament. If the glass heats, it expands, and as it cools, it contacts. This expansion and contraction could cause the glass to become brittle and crack, reducing the life of the bulbs. Many bulbs are now partially filled with an inert gas. It is also useful to remove the oxygen to reduce the possibility of the filament actually burning. When the original filaments were replaced with more efficient tungsten filaments, atoms from the tungsten would evaporate off the filament at such high temperatures. The atoms collide with the atoms of the inert gas and land back on the filament.

Problems

Exercise:

Problem:

An aluminum wire 1.628 mm in diameter (14-gauge) carries a current of 3.00 amps. (a) What is the absolute value of the charge density in the wire? (b) What is the drift velocity of the electrons? (c) What would be the drift velocity if the same gauge copper were used instead of aluminum? The density of copper is 8.96 g/cm^3 and the density of aluminum is 2.70 g/cm^3. The molar mass of aluminum is 26.98 g/mol and the molar mass of copper is 63.5 g/mol. Assume each atom of metal contributes one free electron.

Exercise:

Problem:

The current of an electron beam has a measured current of $I = 50.00\ \mu\text{A}$ with a radius of 1.00 mm. What is the magnitude of the current density of the beam?

Solution:

$|J| = 15.92 \text{ A/m}^2$

Exercise:

Problem:

A high-energy proton accelerator produces a proton beam with a radius of $r = 0.90 \text{ mm}$. The beam current is $I = 9.00\ \mu\text{A}$ and is constant. The charge density of the beam is $n = 6.00 \times 10^{11}$ protons per cubic meter. (a) What is the current density of the beam? (b) What is the drift velocity of the beam? (c) How much time does it take for 1.00×10^{10} protons to be emitted by the accelerator?

Exercise:

Problem:

Consider a wire of a circular cross-section with a radius of $R = 3.00 \text{ mm}$. The magnitude of the current density is modeled as $J = cr^2 = 5.00 \times 10^6 \frac{\text{A}}{\text{m}^4} r^2$. What is the current through the inner section of the wire from the center to $r = 0.5R$?

Solution:

$I = 3.98 \times 10^{-5} \text{ A}$

Exercise:

Problem:

A cylindrical wire has a current density from the center of the wire's cross section as $J(r) = Cr^2$ where r is in meters, J is in amps per square meter, and $C = 10^3 \text{ A/m}^4$. This current density continues to the end of the wire at a radius of 1.0 mm. Calculate the current just outside of this wire.

Exercise:

Problem:

The current supplied to an air conditioner unit is 4.00 amps. The air conditioner is wired using a 10-gauge (diameter 2.588 mm) wire. The charge density is $n = 8.48 \times 10^{28} \frac{\text{electrons}}{\text{m}^3}$. Find the magnitude of (a) current density and (b) the drift velocity.

Solution:

a. $|J| = 7.60 \times 10^5 \frac{\text{A}}{\text{m}^2}$; b. $v_\text{d} = 5.60 \times 10^{-5} \frac{\text{m}}{\text{s}}$

Glossary

current density
flow of charge through a cross-sectional area divided by the area

drift velocity
velocity of a charge as it moves nearly randomly through a conductor, experiencing multiple collisions, averaged over a length of a conductor, whose magnitude is the length of conductor traveled divided by the time it takes for the charges to travel the length

Resistivity and Resistance

By the end of this section, you will be able to:

- Differentiate between resistance and resistivity
- Define the term conductivity
- Describe the electrical component known as a resistor
- State the relationship between resistance of a resistor and its length, cross-sectional area, and resistivity
- State the relationship between resistivity and temperature

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—that are necessary to maintain a current. All such devices create a potential difference and are referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference *V* that creates an electrical field. The electrical field, in turn, exerts force on free charges, causing current. The amount of current depends not only on the magnitude of the voltage, but also on the characteristics of the material that the current is flowing through. The material can resist the flow of the charges, and the measure of how much a material resists the flow of charges is known as the *resistivity*. This resistivity is crudely analogous to the friction between two materials that resists motion.

Resistivity

When a voltage is applied to a conductor, an electrical field $\vec{\mathbf{E}}$ is created, and charges in the conductor feel a force due to the electrical field. The current density $\vec{\mathbf{J}}$ that results depends on the electrical field and the properties of the material. This dependence can be very complex. In some materials, including metals at a given temperature, the current density is approximately proportional to the electrical field. In these cases, the current density can be modeled as

Equation:

$$\vec{\mathbf{J}} = \sigma\vec{\mathbf{E}},$$

where σ is the **electrical conductivity**. The electrical conductivity is analogous to thermal conductivity and is a measure of a material's ability to conduct or transmit electricity. Conductors have a higher electrical conductivity than insulators. Since the electrical conductivity is $\sigma = J/E$, the units are

Equation:

$$\sigma = \frac{[J]}{[E]} = \frac{\text{A/m}^2}{\text{V/m}} = \frac{\text{A}}{\text{V} \cdot \text{m}}.$$

Here, we define a unit named the **ohm** with the Greek symbol uppercase omega, Ω. The unit is named after Georg Simon Ohm, whom we will discuss later in this chapter. The Ω is used to avoid confusion with the number 0. One ohm equals one volt per amp: $1\,\Omega\ = 1\,\text{V/A}$. The units of electrical conductivity are therefore $(\,\Omega\ \cdot \text{m})^{-1}$.

Conductivity is an intrinsic property of a material. Another intrinsic property of a material is the **resistivity**, or electrical resistivity. The resistivity of a material is a measure of how strongly a material opposes the flow of electrical current. The symbol for resistivity is the lowercase Greek letter rho, ρ, and resistivity is the reciprocal of electrical conductivity:

Equation:

$$\rho = \frac{1}{\sigma}.$$

The unit of resistivity in SI units is the ohm-meter $(\Omega \cdot \mathrm{m})$. We can define the resistivity in terms of the electrical field and the current density,

Note:

Equation:

$$\rho = \frac{E}{J}.$$

The greater the resistivity, the larger the field needed to produce a given current density. The lower the resistivity, the larger the current density produced by a given electrical field. Good conductors have a high conductivity and low resistivity. Good insulators have a low conductivity and a high resistivity. [link] lists resistivity and conductivity values for various materials.

Material	**Conductivity,** σ $(\Omega \cdot \mathrm{m})^{-1}$	**Resistivity,** ρ $(\Omega \cdot \mathrm{m})$	**Temperature Coefficient,** α $(^{\circ}\mathrm{C})^{-1}$
Conductors			
Silver	6.29×10^{7}	1.59×10^{-8}	0.0038
Copper	5.95×10^{7}	1.68×10^{-8}	0.0039
Gold	4.10×10^{7}	2.44×10^{-8}	0.0034
Aluminum	3.77×10^{7}	2.65×10^{-8}	0.0039

Material	**Conductivity, σ** $(\Omega\cdot\mathrm{m})^{-1}$	**Resistivity, ρ** $(\Omega\cdot\mathrm{m})$	**Temperature Coefficient, α** $(^\circ\mathrm{C})^{-1}$
Tungsten	1.79×10^{7}	5.60×10^{-8}	0.0045
Iron	1.03×10^{7}	9.71×10^{-8}	0.0065
Platinum	0.94×10^{7}	10.60×10^{-8}	0.0039
Steel	0.50×10^{7}	20.00×10^{-8}	
Lead	0.45×10^{7}	22.00×10^{-8}	
Manganin (Cu, Mn, Ni alloy)	0.21×10^{7}	48.20×10^{-8}	0.000002
Constantan (Cu, Ni alloy)	0.20×10^{7}	49.00×10^{-8}	0.00003
Mercury	0.10×10^{7}	98.00×10^{-8}	0.0009
Nichrome (Ni, Fe, Cr alloy)	0.10×10^{7}	100.00×10^{-8}	0.0004
Semiconductors[1]			
Carbon (pure)	2.86×10^{4}	3.50×10^{-5}	−0.0005
Carbon	$(2.86-1.67)\times10^{-6}$	$(3.5-60)\times10^{-5}$	−0.0005
Germanium (pure)		600×10^{-3}	−0.048
Germanium		$(1-600)\times10^{-3}$	−0.050
Silicon (pure)		2300	−0.075
Silicon		0.1 – 2300	−0.07
Insulators			
Amber	2.00×10^{-15}	5×10^{14}	
Glass	$10^{-9}-10^{-14}$	$10^{9}-10^{14}$	

Material	**Conductivity,** σ $(\Omega\cdot\mathrm{m})^{-1}$	**Resistivity,** ρ $(\Omega\cdot\mathrm{m})$	**Temperature Coefficient,** α $(^{\circ}\mathrm{C})^{-1}$
Lucite	$<10^{-13}$	$>10^{13}$	
Mica	$10^{-11}-10^{-15}$	$10^{11}-10^{15}$	
Quartz (fused)	1.33×10^{-18}	75×10^{16}	
Rubber (hard)	$10^{-13}-10^{-16}$	$10^{13}-10^{16}$	
Sulfur	10^{-15}	10^{15}	
Teflon$^{\mathrm{TM}}$	$<10^{-13}$	$>10^{13}$	
Wood	$10^{-8}-10^{-11}$	$10^{8}-10^{11}$	

Resistivities and Conductivities of Various Materials at 20 °C[1] Values depend strongly on amounts and types of impurities.

The materials listed in the table are separated into categories of conductors, semiconductors, and insulators, based on broad groupings of resistivity. Conductors have the smallest resistivity, and insulators have the largest; semiconductors have intermediate resistivity. Conductors have varying but large, free charge densities, whereas most charges in insulators are bound to atoms and are not free to move. Semiconductors are intermediate, having far fewer free charges than conductors, but having properties that make the number of free charges depend strongly on the type and amount of impurities in the semiconductor. These unique properties of semiconductors are put to use in modern electronics, as we will explore in later chapters.

Note:
Exercise:

Problem:

Check Your Understanding Copper wires use routinely used for extension cords and house wiring for several reasons. Copper has the highest electrical conductivity rating, and therefore the lowest resistivity rating, of all nonprecious metals. Also important is the tensile strength, where the tensile strength is a measure of the force required to pull an object to the point where it breaks. The tensile strength of a material is the maximum amount of tensile stress it can take before breaking. Copper has a high tensile strength, $2 \times 10^8 \frac{\text{N}}{\text{m}^2}$. A third important characteristic is ductility. Ductility is a measure of a material's ability to be drawn into wires and a measure of the flexibility of the material, and copper has a high ductility. Summarizing, for a conductor to be a suitable candidate for making wire, there are at least three important characteristics: low resistivity, high tensile strength, and high ductility. What other materials are used for wiring and what are the advantages and disadvantages?

Solution:

Silver, gold, and aluminum are all used for making wires. All four materials have a high conductivity, silver having the highest. All four can easily be drawn into wires and have a high tensile strength, though not as high as copper. The obvious disadvantage of gold and silver is the cost, but silver and gold wires are used for special applications, such as speaker wires. Gold does not oxidize, making better connections between components. Aluminum wires do have their drawbacks. Aluminum has a higher resistivity than copper, so a larger diameter is needed to match the resistance per length of copper wires, but aluminum is cheaper than copper, so this is not a major drawback. Aluminum wires do not have as high of a ductility and tensile strength as copper, but the ductility and tensile strength is within acceptable levels. There are a few concerns that must be addressed in using aluminum and care must be used when making connections. Aluminum has a higher rate of thermal expansion than copper, which can lead to loose connections and a possible fire hazard. The oxidation of aluminum does not conduct and can cause problems. Special techniques must be used when using aluminum wires and components, such as electrical outlets, must be designed to accept aluminum wires.

Note:
View this interactive simulation to see what the effects of the cross-sectional area, the length, and the resistivity of a wire are on the resistance of a conductor. Adjust the variables using slide bars and see if the resistance becomes smaller or larger.

Temperature Dependence of Resistivity

Looking back at [link], you will see a column labeled "Temperature Coefficient." The resistivity of some materials has a strong temperature dependence. In some materials, such as copper, the resistivity increases with increasing temperature. In fact, in most conducting metals, the resistivity increases with increasing temperature. The increasing temperature causes increased vibrations of the atoms in the lattice structure of the metals, which impede the motion of the electrons. In other materials, such as carbon, the resistivity decreases with increasing temperature. In many materials, the dependence is approximately linear and can be modeled using a linear equation:

Note:
Equation:

$$\rho \approx \rho_0 \left[1 + \alpha \left(T - T_0\right)\right],$$

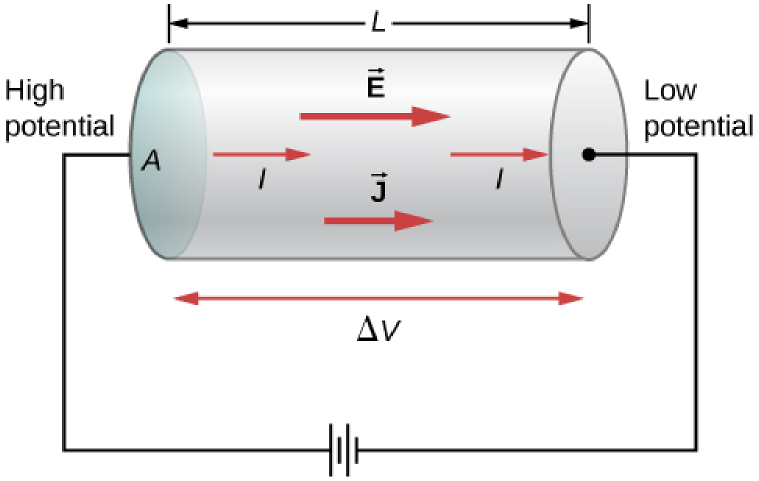
where ρ is the resistivity of the material at temperature T, α is the temperature coefficient of the material, and ρ_0 is the resistivity at T_0, usually taken as $T_0 = 20.00\ ^\circ\text{C}$.

Note also that the temperature coefficient α is negative for the semiconductors listed in [link], meaning that their resistivity decreases with increasing temperature. They become better conductors at higher temperature, because increased thermal agitation increases the number of free charges available to carry current. This property of decreasing ρ with temperature is also related to the type and amount of impurities present in the semiconductors.

Resistance

We now consider the resistance of a wire or component. The resistance is a measure of how difficult it is to pass current through a wire or component. Resistance depends on the resistivity. The resistivity is a characteristic of the material used to fabricate a wire or other electrical component, whereas the resistance is a characteristic of the wire or component.

To calculate the resistance, consider a section of conducting wire with cross-sectional area A, length L, and resistivity ρ. A battery is connected across the conductor, providing a potential difference ΔV across it ([link]). The potential difference produces an electrical field that is proportional to the current density, according to $\vec{\mathbf{E}} = \rho\vec{\mathbf{J}}$.



A potential provided by a battery is applied to a segment of a conductor with a cross-sectional area A and a length L.

The magnitude of the electrical field across the segment of the conductor is equal to the voltage divided by the length, $E = V/L$, and the magnitude of the current density is equal to the current divided by the cross-sectional area, $J = I/A$. Using this information and recalling that the electrical field is proportional to the resistivity and the current density, we can see that the voltage is proportional to the current:

Equation:

$$\begin{aligned} E &= \rho J \\ \frac{V}{L} &= \rho \frac{I}{A} \\ V &= \left(\rho \frac{L}{A}\right) I. \end{aligned}$$

Note:

Resistance

The ratio of the voltage to the current is defined as the **resistance** R:

Equation:

$$R \equiv \frac{V}{I}.$$

The resistance of a cylindrical segment of a conductor is equal to the resistivity of the material times the length divided by the area:

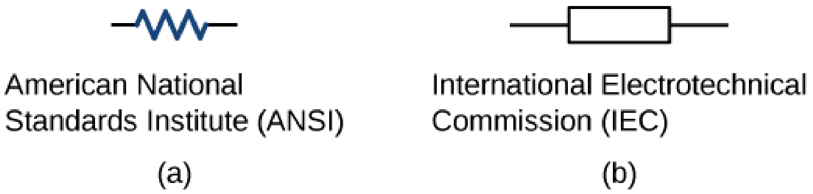
Equation:

$$R \equiv \frac{V}{I} = \rho \frac{L}{A}.$$

The unit of resistance is the ohm, Ω. For a given voltage, the higher the resistance, the lower the current.

Resistors

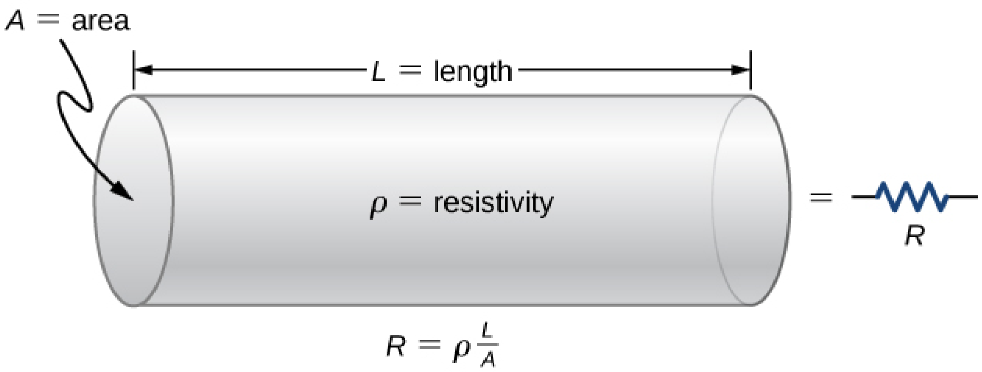
A common component in electronic circuits is the resistor. The resistor can be used to reduce current flow or provide a voltage drop. [link] shows the symbols used for a resistor in schematic diagrams of a circuit. Two commonly used standards for circuit diagrams are provided by the American National Standard Institute (ANSI, pronounced “AN-see”) and the International Electrotechnical Commission (IEC). Both systems are commonly used. We use the ANSI standard in this text for its visual recognition, but we note that for larger, more complex circuits, the IEC standard may have a cleaner presentation, making it easier to read.


American National Standards Institute (ANSI)
(a)
International Electrotechnical Commission (IEC)
(b)

Symbols for a resistor used in circuit diagrams. (a) The ANSI symbol; (b) the IEC symbol.

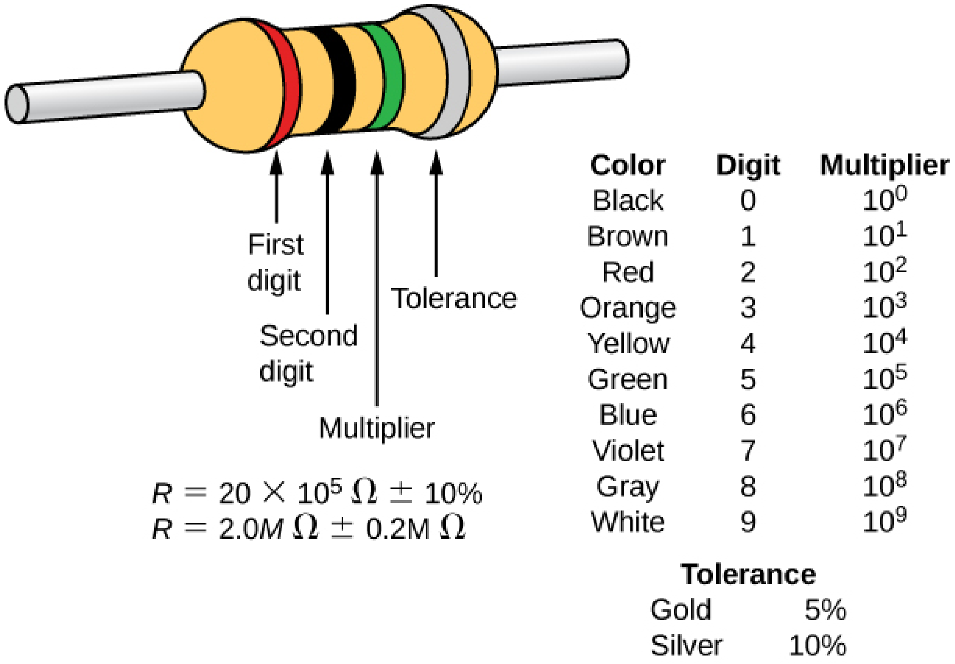
Material and shape dependence of resistance

A resistor can be modeled as a cylinder with a cross-sectional area A and a length L, made of a material with a resistivity ρ ([link]). The resistance of the resistor is $R=\rho\frac{L}{A}$.


A = area
L = length
ρ = resistivity
=
R
$R = \rho \frac{L}{A}$

A model of a resistor as a uniform cylinder of length *L* and cross-sectional area *A*. Its resistance to the flow of current is analogous to the resistance posed by a pipe to fluid flow. The longer the cylinder, the greater its resistance. The larger its cross-sectional area *A*, the smaller its resistance.

The most common material used to make a resistor is carbon. A carbon track is wrapped around a ceramic core, and two copper leads are attached. A second type of resistor is the metal film resistor, which also has a ceramic core. The track is made from a metal oxide material, which has semiconductive properties similar to carbon. Again, copper leads are inserted into the ends of the resistor. The resistor is then painted and marked for identification. A resistor has four colored bands, as shown in [link].


First digit
Second digit
Multiplier
Tolerance
$R = 20 \times 10^5\ \Omega \pm 10\%$
$R = 2.0M\ \Omega \pm 0.2M\ \Omega$

Color	Digit	Multiplier
Black	0	10^0
Brown	1	10^1
Red	2	10^2
Orange	3	10^3
Yellow	4	10^4
Green	5	10^5
Blue	6	10^6
Violet	7	10^7
Gray	8	10^8
White	9	10^9

Tolerance	
Gold	5%
Silver	10%

Many resistors resemble the figure shown above. The four bands are used to identify the resistor. The first two colored bands represent the first two digits of the resistance of the resistor. The third color is the multiplier. The fourth color represents the tolerance of the resistor. The resistor shown has a resistance of $20 \times 10^5\,\Omega \pm 10\%$.

Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of $10^{12}\,\Omega$ or more. A dry person may have a hand-to-foot resistance of $10^5\,\Omega$, whereas the resistance of the human heart is about $10^3\,\Omega$. A meter-long piece of large-diameter copper wire may have a resistance of $10^{-5}\,\Omega$, and superconductors have no resistance at all at low temperatures. As we have seen, resistance is related to the shape of an object and the material of which it is composed.

Example:

Current Density, Resistance, and Electrical field for a Current-Carrying Wire

Calculate the current density, resistance, and electrical field of a 5-m length of copper wire with a diameter of 2.053 mm (12-gauge) carrying a current of $I = 10\,\text{mA}$.

Strategy

We can calculate the current density by first finding the cross-sectional area of the wire, which is $A = 3.31\,\text{mm}^2$, and the definition of current density $J = \frac{I}{A}$. The resistance can be found using the length of the wire $L = 5.00\,\text{m}$, the area, and the resistivity of copper $\rho = 1.68 \times 10^{-8}\,\Omega \cdot \text{m}$, where $R = \rho\frac{L}{A}$. The resistivity and current density can be used to find the electrical field.

Solution

First, we calculate the current density:

Equation:

$$J = \frac{I}{A} = \frac{10 \times 10^{-3}\text{A}}{3.31 \times 10^{-6}\text{m}^2} = 3.02 \times 10^3\,\frac{\text{A}}{\text{m}^2}.$$

The resistance of the wire is

Equation:

$$R = \rho\frac{L}{A} = \left(1.68 \times 10^{-8}\,\Omega \cdot \text{m}\right)\frac{5.00\,\text{m}}{3.31 \times 10^{-6}\text{m}^2} = 0.025\,\Omega\,.$$

Finally, we can find the electrical field:

Equation:

$$E = \rho J = 1.68 \times 10^{-8}\,\Omega \cdot \text{m}\left(3.02 \times 10^3\,\frac{\text{A}}{\text{m}^2}\right) = 5.07 \times 10^{-5}\,\frac{\text{V}}{\text{m}}.$$

Significance

From these results, it is not surprising that copper is used for wires for carrying current because the resistance is quite small. Note that the current density and electrical field are independent of the length of the wire, but the voltage depends on the length.

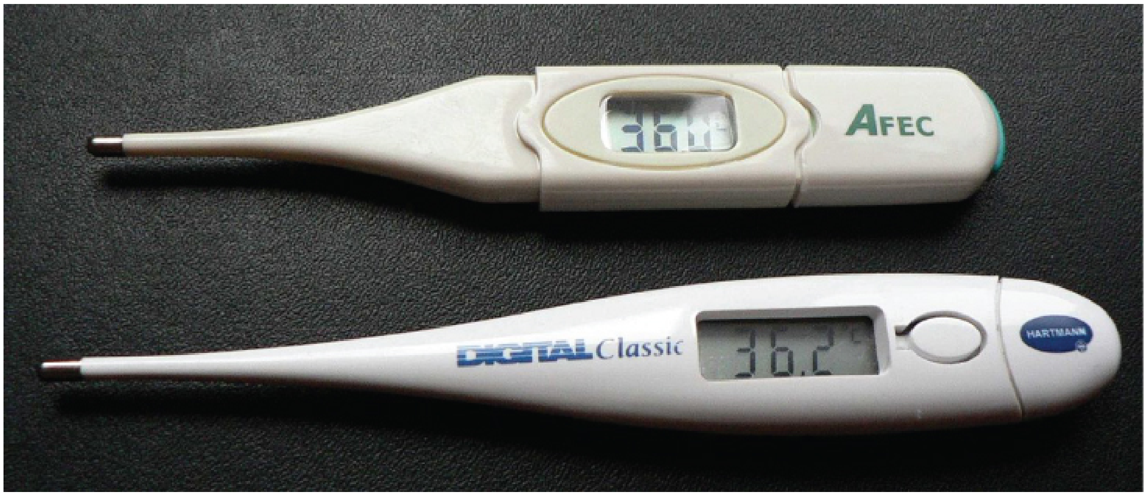
The resistance of an object also depends on temperature, since R_0 is directly proportional to ρ. For a cylinder, we know $R = \rho \frac{L}{A}$, so if L and A do not change greatly with temperature, R has the same temperature dependence as ρ. (Examination of the coefficients of linear expansion shows them to be about two orders of magnitude less than typical temperature coefficients of resistivity, so the effect of temperature on L and A is about two orders of magnitude less than on ρ.) Thus,

Note:
Equation:

$$R = R_0(1 + \alpha \Delta T)$$

is the temperature dependence of the resistance of an object, where R_0 is the original resistance (usually taken to be $20.00\ ^\circ\text{C}$) and R is the resistance after a temperature change ΔT. The color code gives the resistance of the resistor at a temperature of $T = 20.00\ ^\circ\text{C}$.

Numerous thermometers are based on the effect of temperature on resistance ([link]). One of the most common thermometers is based on the thermistor, a semiconductor crystal with a strong temperature dependence, the resistance of which is measured to obtain its temperature. The device is small, so that it quickly comes into thermal equilibrium with the part of a person it touches.



These familiar thermometers are based on the automated measurement of a thermistor’s temperature-dependent resistance.

Example:

Calculating Resistance

Although caution must be used in applying $\rho = \rho_0(1+\alpha\Delta T)$ and $R = R_0(1+\alpha\Delta T)$ for temperature changes greater than $100\ ^\circ\text{C}$, for tungsten, the equations work reasonably well for very large temperature changes. A tungsten filament at $20\ ^\circ\text{C}$ has a resistance of $0.350\,\Omega$. What would the resistance be if the temperature is increased to $2850\ ^\circ\text{C}$?

Strategy

This is a straightforward application of $R = R_0(1+\alpha\Delta T)$, since the original resistance of the filament is given as $R_0 = 0.350\,\Omega$ and the temperature change is $\Delta T = 2830\ ^\circ\text{C}$.

Solution

The resistance of the hotter filament *R* is obtained by entering known values into the above equation:

Equation:

$$R = R_0\left(1+\alpha\Delta T\right) = \ (0.350\,\Omega)\left[1+\left(\frac{4.5\ \times\ 10^{-3}}{^\circ\text{C}}\right)(2830\ ^\circ\text{C})\right] = 4.8\,\Omega\,.$$

Significance

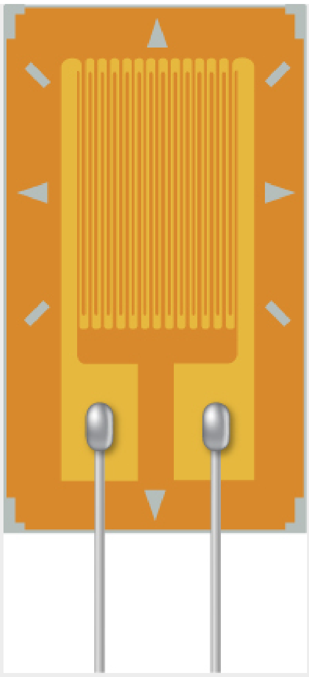
Notice that the resistance changes by more than a factor of 10 as the filament warms to the high temperature and the current through the filament depends on the resistance of the filament and the voltage applied. If the filament is used in an incandescent light bulb, the initial current through the filament when the bulb is first energized will be higher than the current after the filament reaches the operating temperature.

Note:

Exercise:

Problem:

Check Your Understanding A strain gauge is an electrical device to measure strain, as shown below. It consists of a flexible, insulating backing that supports a conduction foil pattern. The resistance of the foil changes as the backing is stretched. How does the strain gauge resistance change? Is the strain gauge affected by temperature changes?



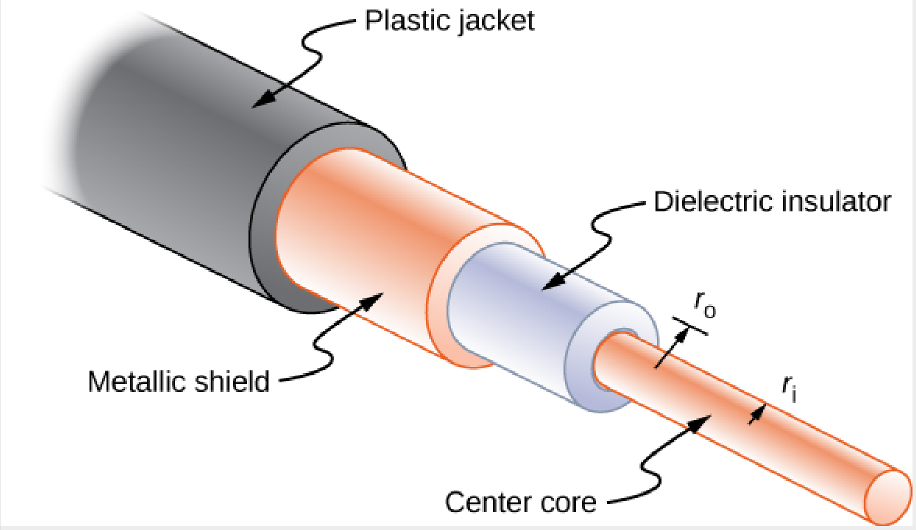
Solution:

The foil pattern stretches as the backing stretches, and the foil tracks become longer and thinner. Since the resistance is calculated as $R = \rho \frac{L}{A}$, the resistance increases as the foil tracks are stretched. When the temperature changes, so does the resistivity of the foil tracks, changing the resistance. One way to combat this is to use two strain gauges, one used as a reference and the other used to measure the strain. The two strain gauges are kept at a constant temperature

Example:

The Resistance of Coaxial Cable

Long cables can sometimes act like antennas, picking up electronic noise, which are signals from other equipment and appliances. Coaxial cables are used for many applications that require this noise to be eliminated. For example, they can be found in the home in cable TV connections or other audiovisual connections. Coaxial cables consist of an inner conductor of radius r_{i} surrounded by a second, outer concentric conductor with radius r_{o} ([link]). The space between the two is normally filled with an insulator such as polyethylene plastic. A small amount of radial leakage current occurs between the two conductors. Determine the resistance of a coaxial cable of length L.



Coaxial cables consist of two concentric conductors separated by insulation. They are often used in cable TV or other audiovisual connections.

Strategy

We cannot use the equation $R=\rho \frac{L}{A}$ directly. Instead, we look at concentric cylindrical shells, with thickness *dr*, and integrate.

Solution

We first find an expression for *dR* and then integrate from r_i to r_o,

Equation:

$$\begin{aligned} dR &= \frac{\rho}{A}dr=\frac{\rho}{2\pi rL}dr, \\ R &= \int_{r_\text{i}}^{r_\text{o}} dR=\int_{r_\text{i}}^{r_\text{o}} \frac{\rho}{2\pi rL}dr=\frac{\rho}{2\pi L}\int_{r_\text{i}}^{r_\text{o}} \frac{1}{r}dr=\frac{\rho}{2\pi L}\ln\frac{r_\text{o}}{r_\text{i}}. \end{aligned}$$

Significance

The resistance of a coaxial cable depends on its length, the inner and outer radii, and the resistivity of the material separating the two conductors. Since this resistance is not infinite, a small leakage current occurs between the two conductors. This leakage current leads to the attenuation (or weakening) of the signal being sent through the cable.

Note:

Exercise:

Problem:

Check Your Understanding The resistance between the two conductors of a coaxial cable depends on the resistivity of the material separating the two conductors, the length of the cable and the inner and outer radius of the two conductor. If you are designing a coaxial cable, how does the resistance between the two conductors depend on these variables?

Solution:

The longer the length, the smaller the resistance. The greater the resistivity, the higher the resistance. The larger the difference between the outer radius and the inner radius, that is, the greater the ratio between the two, the greater the resistance. If you are attempting to maximize the resistance, the choice of the values for these variables will depend on the application. For example, if the cable must be flexible, the choice of materials may be limited.

Note:
View this simulation to see how the voltage applied and the resistance of the material the current flows through affects the current through the material. You can visualize the collisions of the electrons and the atoms of the material effect the temperature of the material.

Summary

- Resistance has units of ohms (Ω), related to volts and amperes by $1\,\Omega\ =\ 1\,\text{V}/\text{A}$.
- The resistance R of a cylinder of length L and cross-sectional area A is $R=\frac{\rho L}{A}$, where ρ is the resistivity of the material.
- Values of ρ in [link] show that materials fall into three groups—conductors, semiconductors, and insulators.
- Temperature affects resistivity; for relatively small temperature changes ΔT, resistivity is $\rho=\rho_0\left(1+\alpha\Delta T\right)$, where ρ_0 is the original resistivity and α is the temperature coefficient of resistivity.
- The resistance R of an object also varies with temperature: $R=R_0\left(1+\alpha\Delta T\right)$, where R_0 is the original resistance, and R is the resistance after the temperature change.

Conceptual Questions

Exercise:

Problem:

The IR drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.

Exercise:

Problem:

Do impurities in semiconducting materials listed in [link] supply free charges? (*Hint*: Examine the range of resistivity for each and determine whether the pure semiconductor has the higher or lower conductivity.)

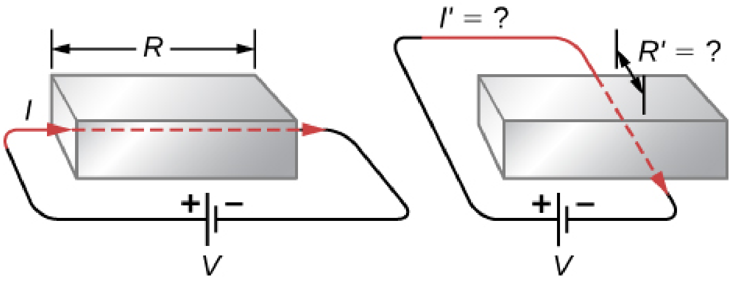
Solution:

In carbon, resistivity increases with the amount of impurities, meaning fewer free charges. In silicon and germanium, impurities decrease resistivity, meaning more free electrons.

Exercise:

Problem:

Does the resistance of an object depend on the path current takes through it? Consider, for example, a rectangular bar—is its resistance the same along its length as across its width?



Exercise:

Problem:

If aluminum and copper wires of the same length have the same resistance, which has the larger diameter? Why?

Solution:

Copper has a lower resistivity than aluminum, so if length is the same, copper must have the smaller diameter.

Problems

Exercise:

Problem:

What current flows through the bulb of a 3.00-V flashlight when its hot resistance is $3.60\ \Omega$?

Exercise:

Problem:

Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

Solution:

$R=6.750\,\mathrm{k}\,\Omega$

Exercise:

Problem:

How many volts are supplied to operate an indicator light on a DVD player that has a resistance of $140\ \Omega$, given that 25.0 mA passes through it?

Exercise:

Problem:

What is the resistance of a 20.0-m-long piece of 12-gauge copper wire having a 2.053-mm diameter?

Solution:

$R=0.10\,\Omega$

Exercise:

Problem:

The diameter of 0-gauge copper wire is 8.252 mm. Find the resistance of a 1.00-km length of such wire used for power transmission.

Exercise:

Problem:

If the 0.100-mm-diameter tungsten filament in a light bulb is to have a resistance of $0.200\ \Omega$ at $20.0\ ^{\circ}\mathrm{C}$, how long should it be?

Solution:

$R=\rho\frac{L}{A}$

$L=3\,\mathrm{cm}$

Exercise:

Problem:

A lead rod has a length of 30.00 cm and a resistance of $5.00\ \mu\Omega$. What is the radius of the rod?

Exercise:

Problem:

Find the ratio of the diameter of aluminum to copper wire, if they have the same resistance per unit length (as they might in household wiring).

Solution:

$$\frac{{}^{R_{\text{Al}}}/_{L_{\text{Al}}}}{{}^{R_{\text{Cu}}}/_{L_{\text{Cu}}}} = \frac{\rho_{\text{Al}}\frac{1}{\pi\left(\frac{D_{\text{Al}}}{2}\right)^2}}{\rho_{\text{Cu}}\frac{1}{\pi\left(\frac{D_{\text{Cu}}}{2}\right)^2}} = \frac{\rho_{\text{Al}}}{\rho_{\text{Cu}}}\left(\frac{D_{\text{Cu}}}{D_{\text{Al}}}\right)^2 = 1, \quad \frac{D_{\text{Al}}}{D_{\text{Cu}}} = \sqrt{\frac{\rho_{\text{Al}}}{\rho_{Cu}}}$$

Exercise:

Problem:

What current flows through a 2.54-cm-diameter rod of pure silicon that is 20.0 cm long, when 1.00×10^3 V is applied to it? (Such a rod may be used to make nuclear-particle detectors, for example.)

Exercise:

Problem:

(a) To what temperature must you raise a copper wire, originally at $20.0\ ^\circ\text{C}$, to double its resistance, neglecting any changes in dimensions? (b) Does this happen in household wiring under ordinary circumstances?

Solution:

a. $R = R_0\left(1 + \alpha\Delta T\right), \quad 2 = 1 + \alpha\Delta T, \quad \Delta T = 256.4\ ^\circ\text{C}, \quad T = 276.4\ ^\circ\text{C};$
b. Under normal conditions, no it should not occur.

Exercise:

Problem:

A resistor made of nichrome wire is used in an application where its resistance cannot change more than 1.00% from its value at $20.0\ ^\circ\text{C}$. Over what temperature range can it be used?

Exercise:

Problem:

Of what material is a resistor made if its resistance is 40.0% greater at $100.0\ ^\circ\text{C}$ than at $20.0\ ^\circ\text{C}$?

Solution:

$$\begin{aligned} R &= \quad R_0\left(1 + \alpha\Delta T\right) \\ \alpha &= 0.006\ ^\circ\text{C}^{-1} \end{aligned}$$, iron

Exercise:

Problem:

An electronic device designed to operate at any temperature in the range from $-10.0\ ^\circ\mathrm{C}$ to $55.0\ ^\circ\mathrm{C}$ contains pure carbon resistors. By what factor does their resistance increase over this range?

Exercise:

Problem:

(a) Of what material is a wire made, if it is 25.0 m long with a diameter of 0.100 mm and has a resistance of $77.7\ \Omega$ at $20.0\ ^\circ\mathrm{C}$? (b) What is its resistance at $150.0\ ^\circ\mathrm{C}$?

Solution:

a. $R = \rho\frac{L}{A}, \quad \rho = 2.44\ \times\ 10^{-8}\ \Omega\ \cdot \mathrm{m}$, gold;
$R = \rho\frac{L}{A}(1 + \alpha\Delta T)$

b. $$R = 2.44\ \times\ 10^{-8}\ \Omega\ \cdot \mathrm{m}\left(\frac{25\ \mathrm{m}}{\pi\left(\frac{0.100 \times 10^{-3}\mathrm{m}}{2}\right)^2}\right)\left(1 + 0.0034\ ^\circ\mathrm{C}^{-1}\left(150\ ^\circ\mathrm{C} - 20\ ^\circ\mathrm{C}\right)\right)$$

$$R = 112\ \Omega$$

Exercise:

Problem:

Assuming a constant temperature coefficient of resistivity, what is the maximum percent decrease in the resistance of a constantan wire starting at $20.0\ ^\circ\mathrm{C}$?

Exercise:

Problem:

A copper wire has a resistance of $0.500\ \Omega$ at $20.0\ ^\circ\mathrm{C}$, and an iron wire has a resistance of $0.525\ \Omega$ at the same temperature. At what temperature are their resistances equal?

Solution:

$$R_{\mathrm{Fe}} = 0.525\ \Omega\,, \quad R_{\mathrm{Cu}} = 0.500\ \Omega\,, \quad \alpha_{\mathrm{Fe}} = 0.0065\ ^\circ\mathrm{C}^{-1} \quad \alpha_{\mathrm{Cu}} = 0.0039\ ^\circ\mathrm{C}^{-1}$$
$$R_{\mathrm{Fe}} = R_{\mathrm{Cu}}$$
$$R_{0\,\mathrm{Fe}}\left(1 + \alpha_{\mathrm{Fe}}\left(T - T_0\right)\right) = R_{0\,\mathrm{Cu}}\left(1 + \alpha_{\mathrm{Cu}}\left(T - T_0\right)\right)$$
$$\frac{R_{0\,\mathrm{Fe}}}{R_{0\,\mathrm{Cu}}}\left(1 + \alpha_{\mathrm{Fe}}\left(T - T_0\right)\right) = 1 + \alpha_{\mathrm{Cu}}\left(T - T_0\right)$$
$$T = 2.91\ ^\circ\mathrm{C}$$

Glossary

electrical conductivity

measure of a material’s ability to conduct or transmit electricity

ohm

(Ω) unit of electrical resistance, $1\,\Omega\ = 1\,\text{V/A}$

resistance

electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, $R = V/I$

resistivity

intrinsic property of a material, independent of its shape or size, directly proportional to the resistance, denoted by ρ

Ohm's Law
By the end of this section, you will be able to:

- Describe Ohm’s law
- Recognize when Ohm’s law applies and when it does not

We have been discussing three electrical properties so far in this chapter: current, voltage, and resistance. It turns out that many materials exhibit a simple relationship among the values for these properties, known as Ohm’s law. Many other materials do not show this relationship, so despite being called Ohm’s law, it is not considered a law of nature, like Newton’s laws or the laws of thermodynamics. But it is very useful for calculations involving materials that do obey Ohm’s law.

Description of Ohm’s Law

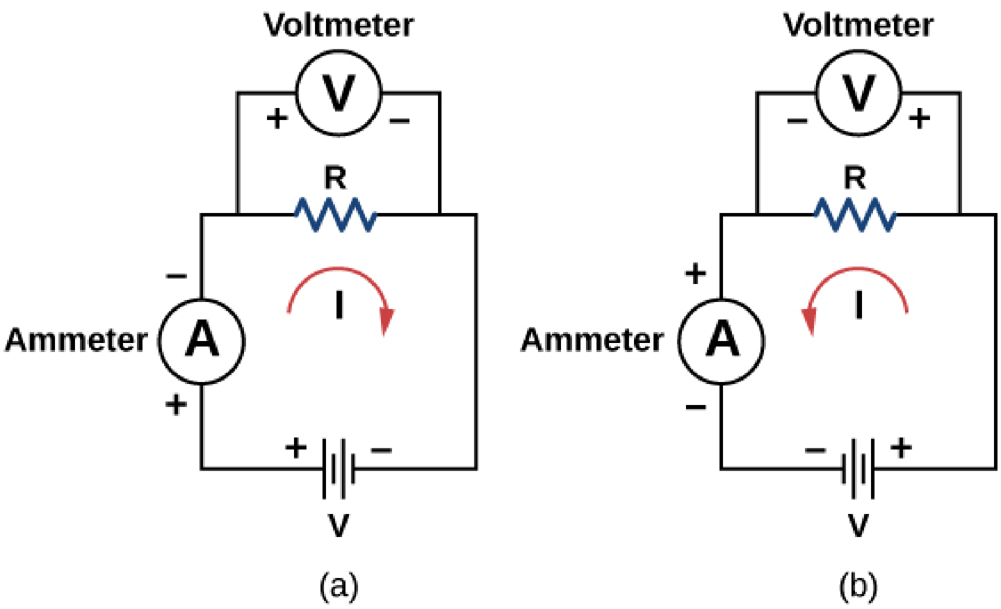
The current that flows through most substances is directly proportional to the voltage V applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is *directly proportional to the voltage applied*:
Equation:

$$I \propto V.$$

This important relationship is the basis for **Ohm’s law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law, which is to say that it is an experimentally observed phenomenon, like friction. Such a linear relationship doesn’t always occur. Any material, component, or device that obeys Ohm’s law, where the current through the device is proportional to the voltage applied, is known as an **ohmic** material or ohmic component. Any material or component that does not obey Ohm’s law is known as a **nonohmic** material or nonohmic component.

Ohm’s Experiment

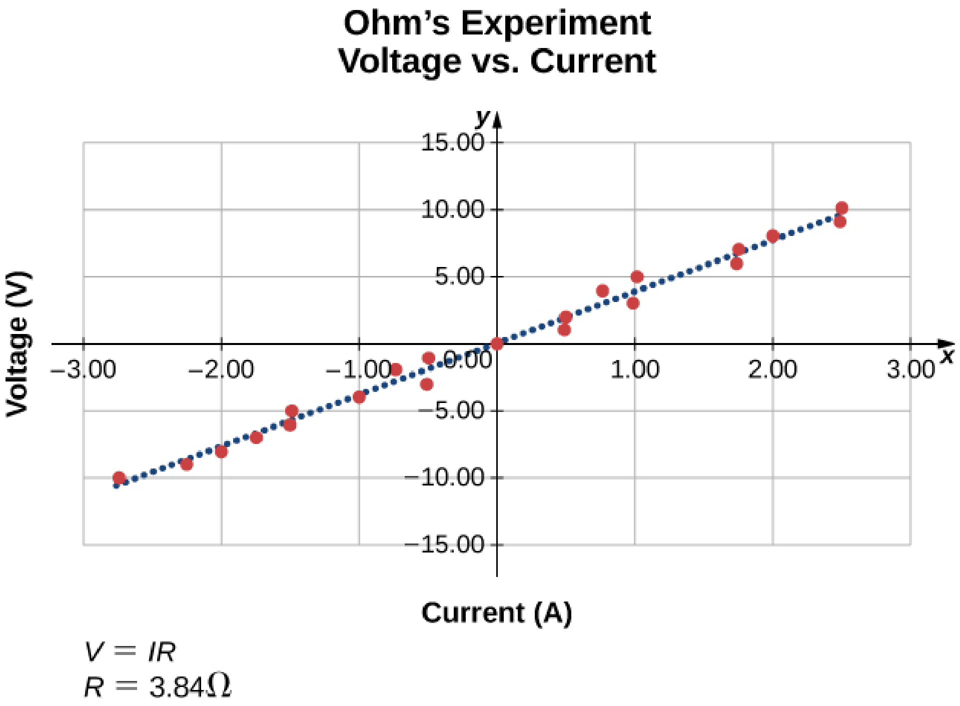
In a paper published in 1827, Georg Ohm described an experiment in which he measured voltage across and current through various simple electrical circuits containing various lengths of wire. A similar experiment is shown in [link]. This experiment is used to observe the current through a resistor that results from an applied voltage. In this simple circuit, a resistor is connected in series with a battery. The voltage is measured with a voltmeter, which must be placed across the resistor (in parallel with the resistor). The current is measured with an ammeter, which must be in line with the resistor (in series with the resistor).



The experimental set-up used to determine if a resistor is an ohmic or nonohmic device. (a) When the battery is attached, the current flows in the clockwise direction and the voltmeter and ammeter have positive readings. (b) When the leads of the battery are switched, the current flows in the counterclockwise direction and the voltmeter and ammeter have negative readings.

In this updated version of Ohm's original experiment, several measurements of the current were made for several different voltages. When the battery was hooked up as in [link](a), the current flowed in the clockwise direction and the readings of the voltmeter and ammeter were positive. Does the behavior of the current change if the current flowed in the opposite direction? To get the current to flow in the opposite direction, the leads of the battery can be switched. When the leads of the battery were switched, the readings of the voltmeter and ammeter readings were negative because the current flowed in the opposite direction, in this case, counterclockwise. Results of a similar experiment are shown in [link].

I(A)	V(V)
–2.74	–10.00
–2.25	–9.00
–2.00	–8.00
–1.75	–7.00
–1.50	–6.00
–1.49	–5.00
–1.00	–4.00
–0.51	–3.00
–0.74	–2.00
–0.49	–1.00
+0.00	+0.00
+0.49	+1.00
+0.50	+2.00
+0.99	+3.00
+0.76	+4.00
+1.01	+5.00
+1.74	+6.00
+1.75	+7.00
+2.00	+8.00
+2.49	+9.00
+2.50	+10.00



A resistor is placed in a circuit with a battery. The voltage applied varies from −10.00 V to +10.00 V, increased by 1.00-V increments. A plot shows values of the voltage versus the current typical of what a casual experimenter might find.

In this experiment, the voltage applied across the resistor varies from −10.00 to +10.00 V, by increments of 1.00 V. The current through the resistor and the voltage across the resistor are measured. A plot is made of the voltage versus the current, and the result is approximately linear. The slope of the line is the resistance, or the voltage divided by the current. This result is known as Ohm’s law:

Note:
Equation:

$$V = IR,$$

where *V* is the voltage measured in volts across the object in question, *I* is the current measured through the object in amps, and *R* is the resistance in units of ohms. As stated previously, any device that shows a linear relationship between the voltage and the current is known as an ohmic device. A resistor is therefore an ohmic device.

Example:

Measuring Resistance

A carbon resistor at room temperature $(20\ ^\circ\text{C})$ is attached to a 9.00-V battery and the current measured through the resistor is 3.00 mA. (a) What is the resistance of the resistor measured in ohms? (b) If the temperature of the resistor is increased to $60\ ^\circ\text{C}$ by heating the resistor, what is the current through the resistor?

Strategy

(a) The resistance can be found using Ohm's law. Ohm's law states that $V = IR$, so the resistance can be found using $R = V/I$.

(b) First, the resistance is temperature dependent so the new resistance after the resistor has been heated can be found using $R = R_0\left(1 + \alpha\Delta T\right)$. The current can be found using Ohm's law in the form $I = V/R$.

Solution

a. Using Ohm's law and solving for the resistance yields the resistance at room temperature:
   Equation:

$$R = \frac{V}{I} = \frac{9.00\ \text{V}}{3.00\ \times\ 10^{-3}\ \text{A}} = 3.00\ \times\ 10^{3}\ \Omega\ = 3.00\ \text{k}\ \Omega.$$

b. The resistance at $60\ ^\circ\text{C}$ can be found using $R = R_0\left(1 + \alpha\Delta T\right)$ where the temperature coefficient for carbon is $\alpha = -0.0005$.
   $R = R_0\left(1 + \alpha\Delta T\right) = 3.00\ \times\ 10^{3}\left(1 - 0.0005\left(60\ ^\circ\text{C} - 20\ ^\circ\text{C}\right)\right) = 2.94\ \text{k}\ \Omega$
   .
   The current through the heated resistor is
   Equation:

$$I = \frac{V}{R} = \frac{9.00\ \text{V}}{2.94\ \times\ 10^{3}\ \Omega} = 3.06\ \times\ 10^{-3}\text{A} = 3.06\ \text{mA}.$$

Significance

A change in temperature of $40\ ^\circ\text{C}$ resulted in a 2.00% change in current. This may not seem like a very great change, but changing electrical characteristics can have a strong effect on the circuits. For this reason, many electronic appliances, such as computers, contain fans to remove the heat dissipated by components in the electric circuits.

Note:
Exercise:

Problem:

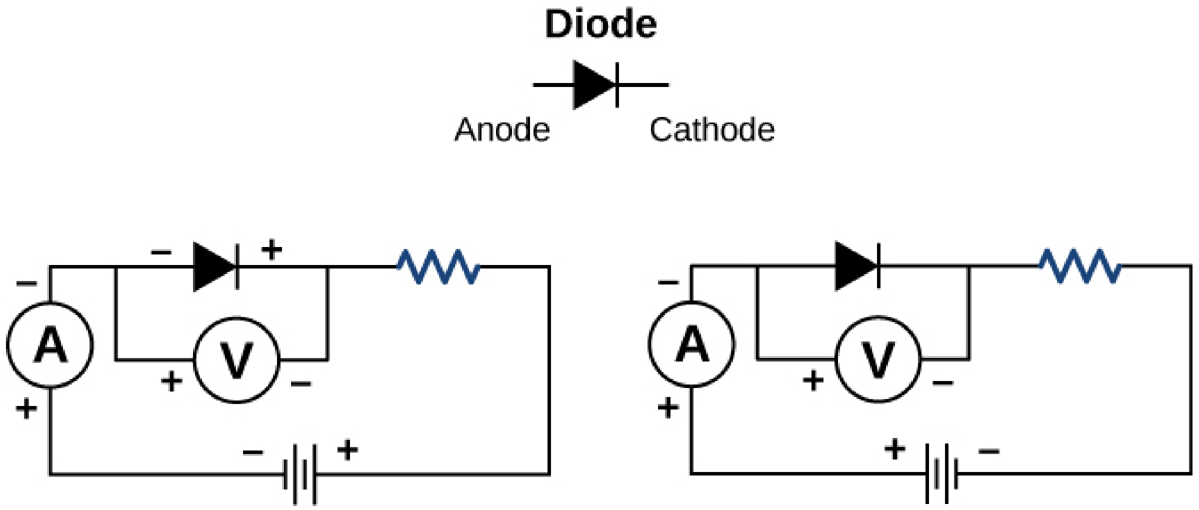
Check Your Understanding The voltage supplied to your house varies as $V\left(t\right)=V_{\text{max}}\,\text{sin}\,\left(2\pi ft\right)$. If a resistor is connected across this voltage, will Ohm's law $V=IR$ still be valid?

Solution:

Yes, Ohm's law is still valid. At every point in time the current is equal to $I(t)=V(t)/R$, so the current is also a function of time, $I\left(t\right)=\frac{V_{\text{max}}}{R}\,\text{sin}\,\left(2\pi ft\right)$.

Note:
See how the equation form of Ohm's law relates to a simple circuit. Adjust the voltage and resistance, and see the current change according to Ohm's law. The sizes of the symbols in the equation change to match the circuit diagram.

Nonohmic devices do not exhibit a linear relationship between the voltage and the current. One such device is the semiconducting circuit element known as a diode. A **diode** is a circuit device that allows current flow in only one direction. A diagram of a simple circuit consisting of a battery, a diode, and a resistor is shown in [link]. Although we do not cover the theory of the diode in this section, the diode can be tested to see if it is an ohmic or a nonohmic device.

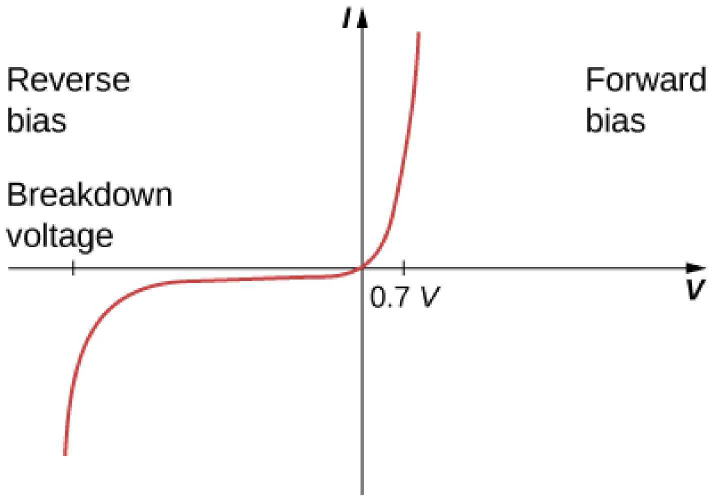


(a) Reverse bias

(b) Forward bias

A diode is a semiconducting device that allows current flow only if the diode is forward biased, which means that the anode is positive and the cathode is negative.

A plot of current versus voltage is shown in [link]. Note that the behavior of the diode is shown as current versus voltage, whereas the resistor operation was shown as voltage versus current. A diode consists of an anode and a cathode. When the anode is at a negative potential and the cathode is at a positive potential, as shown in part (a), the diode is said to have reverse bias. With reverse bias, the diode has an extremely large resistance and there is very little current flow—essentially zero current—through the diode and the resistor. As the voltage applied to the circuit increases, the current remains essentially zero, until the voltage reaches the breakdown voltage and the diode conducts current, as shown in [link]. When the battery and the potential across the diode are reversed, making the anode positive and the cathode negative, the diode conducts and current flows through the diode if the voltage is greater than 0.7 V. The resistance of the diode is close to zero. (This is the reason for the resistor in the circuit; if it were not there, the current would become very large.) You can see from the graph in [link] that the voltage and the current do not have a linear relationship. Thus, the diode is an example of a nonohmic device.



When the voltage across the diode is negative and small, there is very little current flow through the diode. As the voltage reaches the breakdown voltage, the diode conducts. When the voltage across the diode is positive and greater than 0.7 V (the actual voltage value depends on the diode), the diode conducts. As the voltage applied increases, the current through the diode increases, but the voltage across the diode remains approximately 0.7 V.

Ohm’s law is commonly stated as $V=IR$, but originally it was stated as a microscopic view, in terms of the current density, the conductivity, and the electrical field. This microscopic view suggests the proportionality $V\propto I$ comes from the drift velocity of the free electrons in the metal that results from an applied electrical field. As stated earlier, the current density is proportional to the applied electrical field. The reformulation of Ohm’s law is credited to Gustav Kirchhoff, whose name we will see again in the next chapter.

Summary

- Ohm’s law is an empirical relationship for current, voltage, and resistance for some common types of circuit elements, including resistors. It does not apply to other devices, such as diodes.

- One statement of Ohm’s law gives the relationship among current I, voltage V, and resistance R in a simple circuit as $V = IR$.
- Another statement of Ohm’s law, on a microscopic level, is $J = \sigma E$.

Conceptual Questions

Exercise:

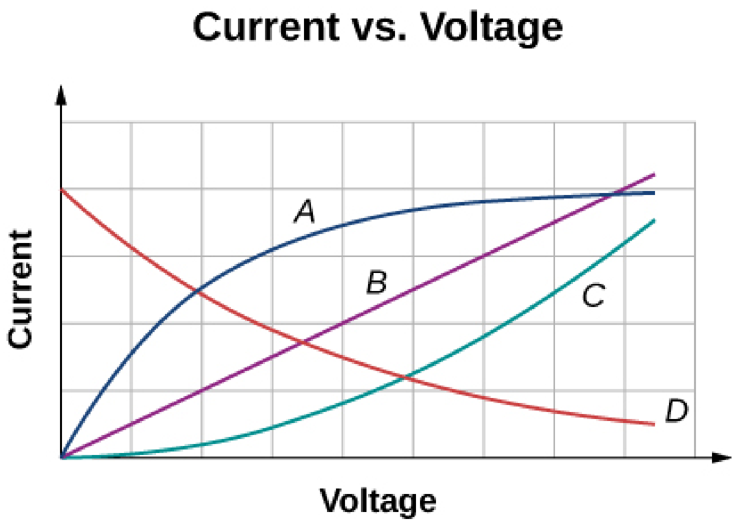
Problem:

In Determining Field from Potential, resistance was defined as $R \equiv \frac{V}{I}$. In this section, we presented Ohm’s law, which is commonly expressed as $V = IR$. The equations look exactly alike. What is the difference between Ohm’s law and the definition of resistance?

Exercise:

Problem:

Shown below are the results of an experiment where four devices were connected across a variable voltage source. The voltage is increased and the current is measured. Which device, if any, is an ohmic device?



Solution:

Device *B* shows a linear relationship and the device is ohmic.

Exercise:

Problem:

The current I is measured through a sample of an ohmic material as a voltage V is applied. (a) What is the current when the voltage is doubled to $2V$ (assume the change in temperature of the material is negligible)? (b) What is the voltage applied is the current measured is $0.2I$ (assume the change in temperature of the material is negligible)? What will happen to the current if the material if the voltage remains constant, but the temperature of the material increases significantly?

Problems

Exercise:

Problem:

A $2.2\text{-k}\,\Omega$ resistor is connected across a D cell battery (1.5 V). What is the current through the resistor?

Exercise:

Problem:

A resistor rated at $250\;\text{k}\,\Omega$ is connected across two D cell batteries (each 1.50 V) in series, with a total voltage of 3.00 V. The manufacturer advertises that their resistors are within 5% of the rated value. What are the possible minimum current and maximum current through the resistor?

Solution:

$$\begin{aligned} R_{\text{min}} &= 2.375\ \times\ 10^5\,\Omega\,,\quad I_{\text{min}} = 12.63\,\mu\,\text{A} \\ R_{\text{max}} &= 2.625\ \times\ 10^5\,\Omega\,,\quad I_{\text{max}} = 11.43\,\mu\,\text{A} \end{aligned}$$

Exercise:

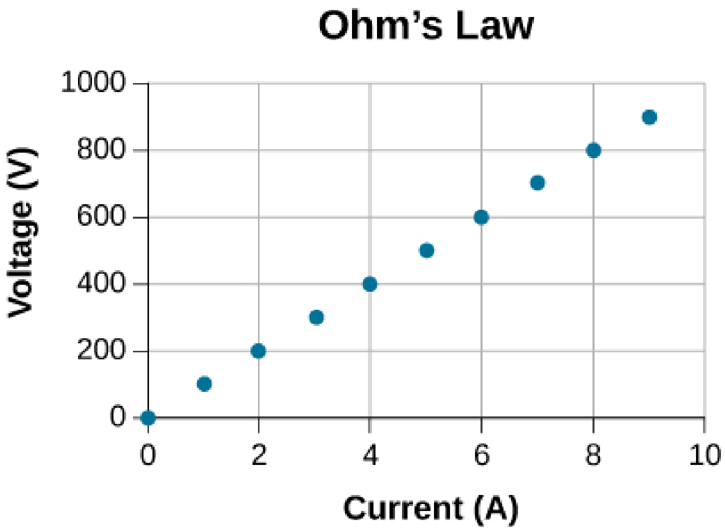
Problem:

A resistor is connected in series with a power supply of 20.00 V. The current measure is 0.50 A. What is the resistance of the resistor?

Exercise:

Problem:

A resistor is placed in a circuit with an adjustable voltage source. The voltage across and the current through the resistor and the measurements are shown below. Estimate the resistance of the resistor.



Solution:

$R = 100\,\Omega$

Exercise:

Problem:

The following table show the measurements of a current through and the voltage across a sample of material. Plot the data, and assuming the object is an ohmic device, estimate the resistance.

I(A)	V(V)
0	3
2	23

***I*(A)**	***V*(V)**
4	39
6	58
8	77
10	100
12	119
14	142
16	162

Glossary

diode	nonohmic circuit device that allows current flow in only one direction
Ohm’s law	empirical relation stating that the current *I* is proportional to the potential difference *V*; it is often written as $V = IR$, where *R* is the resistance
ohmic	type of a material for which Ohm’s law is valid, that is, the voltage drop across the device is equal to the current times the resistance
nonohmic	type of a material for which Ohm’s law is not valid

Electrical Energy and Power

By the end of this section, you will be able to:

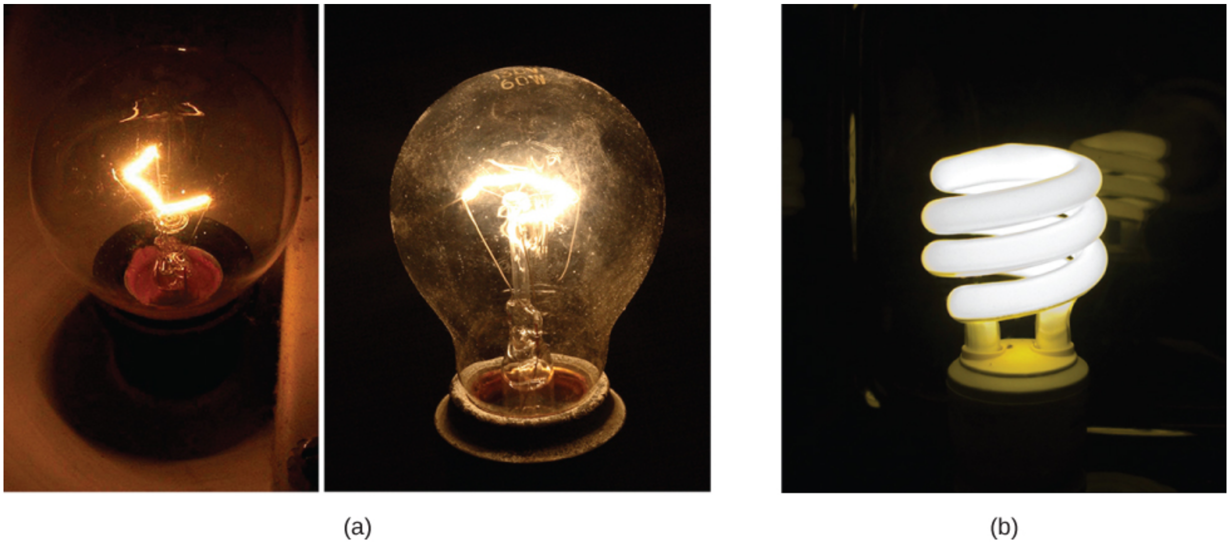
- Express electrical power in terms of the voltage and the current
- Describe the power dissipated by a resistor in an electric circuit
- Calculate the energy efficiency and cost effectiveness of appliances and equipment

In an electric circuit, electrical energy is continuously converted into other forms of energy. For example, when a current flows in a conductor, electrical energy is converted into thermal energy within the conductor. The electrical field, supplied by the voltage source, accelerates the free electrons, increasing their kinetic energy for a short time. This increased kinetic energy is converted into thermal energy through collisions with the ions of the lattice structure of the conductor. In Work and Kinetic Energy, we defined power as the rate at which work is done by a force measured in watts. Power can also be defined as the rate at which energy is transferred. In this section, we discuss the time rate of energy transfer, or power, in an electric circuit.

Power in Electric Circuits

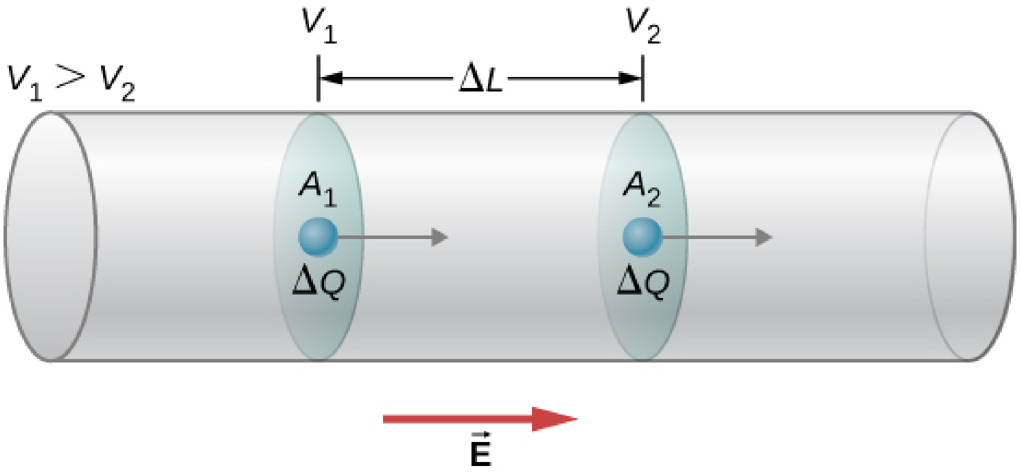
Power is associated by many people with electricity. Power transmission lines might come to mind. We also think of light bulbs in terms of their power ratings in watts. What is the expression for **electric power**?

Let us compare a 25-W bulb with a 60-W bulb ([link](a)). The 60-W bulb glows brighter than the 25-W bulb. Although it is not shown, a 60-W light bulb is also warmer than the 25-W bulb. The heat and light is produced by from the conversion of electrical energy. The kinetic energy lost by the electrons in collisions is converted into the internal energy of the conductor and radiation. How are voltage, current, and resistance related to electric power?



(a) Pictured above are two incandescent bulbs: a 25-W bulb (left) and a 60-W bulb (right). The 60-W bulb provides a higher intensity light than the 25-W bulb. The electrical energy supplied to the light bulbs is converted into heat and light. (b) This compact fluorescent light (CFL) bulb puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit a: modification of works by “Dickbauch”/Wikimedia Commons and Greg Westfall; credit b: modification of work by “dbgg1979”/Flickr)

To calculate electric power, consider a voltage difference existing across a material ([link]). The electric potential V_1 is higher than the electric potential at V_2, and the voltage difference is negative $V = V_2 - V_1$. As discussed in Electric Potential, an electrical field exists between the two potentials, which points from the higher potential to the lower potential. Recall that the electrical potential is defined as the potential energy per charge, $V = \Delta U / q$, and the charge ΔQ loses potential energy moving through the potential difference.



When there is a potential difference across a conductor, an electrical field is present that points in the direction from the higher potential to the lower potential.

If the charge is positive, the charge experiences a force due to the electrical field $\vec{\mathbf{F}} = m\vec{\mathbf{a}} = \Delta Q\vec{\mathbf{E}}$. This force is necessary to keep the charge moving. This force does not act to accelerate the charge through the entire distance ΔL because of the interactions of the charge with atoms and free electrons in the material. The speed, and therefore the kinetic energy, of the charge do not increase during the entire trip across ΔL, and charge passing through area A_2 has the same drift velocity v_{d} as the charge that passes through area A_1. However, work is done on the charge, by the electrical field, which changes the potential energy. Since the change in the electrical potential difference is negative, the electrical field is found to be

Equation:

$$E = -\frac{(V_2 - V_1)}{\Delta L} = \frac{V}{\Delta L}.$$

The work done on the charge is equal to the electric force times the length at which the force is applied,

Equation:

$$W = F\Delta L = (\Delta QE)\Delta L = \left(\Delta Q\frac{V}{\Delta L}\right)\Delta L = \Delta QV = \Delta U.$$

The charge moves at a drift velocity v_{d} so the work done on the charge results in a loss of potential energy, but the average kinetic energy remains constant. The lost electrical potential energy appears as thermal energy in the material. On a microscopic scale, the energy transfer is due to collisions between the charge and the molecules of the material, which leads to an increase in temperature in the material. The loss of potential energy results in an increase in the temperature of the material, which is dissipated as radiation. In a resistor, it is dissipated as heat, and in a light bulb, it is dissipated as heat and light.

The power dissipated by the material as heat and light is equal to the time rate of change of the work:
Equation:

$$P = \frac{\Delta U}{\Delta t} = -\frac{\Delta QV}{\Delta t} = IV.$$

With a resistor, the voltage drop across the resistor is dissipated as heat. Ohm’s law states that the voltage across the resistor is equal to the current times the resistance, $V = IR$. The power dissipated by the resistor is therefore
Equation:

$$P = IV = I\left(IR\right) = I^2R \text{ or } P = IV = \left(\frac{V}{R}\right)V = \frac{V^2}{R}.$$

If a resistor is connected to a battery, the power dissipated as radiant energy by the wires and the resistor is equal to $P = IV = I^2R = \frac{V^2}{R}$. The power supplied from the battery is equal to current times the voltage, $P = IV$.

Note:
Electric Power
The electric power gained or lost by any device has the form
Equation:

$$P = IV.$$

The power dissipated by a resistor has the form
Equation:

$$P = I^2R = \frac{V^2}{R}.$$

Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature, its resistance is higher, too.

Example:

Calculating Power in Electric Devices

A DC winch motor is rated at 20.00 A with a voltage of 115 V. When the motor is running at its maximum power, it can lift an object with a weight of 4900.00 N a distance of 10.00 m, in 30.00 s, at a constant speed. (a) What is the power consumed by the motor? (b) What is the power used in lifting the object? Ignore air resistance. (c) Assuming that the difference in the power consumed by the motor and the power used lifting the object are dissipated as heat by the resistance of the motor, estimate the resistance of the motor?

Strategy

(a) The power consumed by the motor can be found using $P = IV$. (b) The power used in lifting the object at a constant speed can be found using $P = Fv$, where the speed is the distance divided by the time. The upward force supplied by the motor is equal to the weight of the object because the acceleration is zero. (c) The resistance of the motor can be found using $P = I^2R$.

Solution

a. The power consumed by the motor is equal to $P = IV$ and the current is given as 20.00 A and the voltage is 115.00 V:
   Equation:

$$P = IV = (20.00\ \text{A})115.00\ \text{V} = 2300.00\ \text{W}.$$

b. The power used lifting the object is equal to $P = Fv$ where the force is equal to the weight of the object (1960 N) and the magnitude of the velocity is $v = \frac{10.00\,\text{m}}{30.00\,\text{s}} = 0.33\,\frac{\text{m}}{\text{s}}$,
Equation:

$$P = Fv = (4900\,\text{N})0.33\,\text{m/s} = 1633.33\,\text{W}.$$

c. The difference in the power equals $2300.00\,\text{W} - 1633.33\,\text{W} = 666.67\,\text{W}$ and the resistance can be found using $P = I^2R$:
Equation:

$$R = \frac{P}{I^2} = \frac{666.67\,\text{W}}{(20.00\,\text{A})^2} = 1.67\,\Omega\,.$$

Significance
The resistance of the motor is quite small. The resistance of the motor is due to many windings of copper wire. The power dissipated by the motor can be significant since the thermal power dissipated by the motor is proportional to the square of the current $\left(P = I^2R\right)$.

Note:
Exercise:

Problem:

Check Your Understanding Electric motors have a reasonably high efficiency. A 100-hp motor can have an efficiency of 90% and a 1-hp motor can have an efficiency of 80%. Why is it important to use high-performance motors?

Solution:

Even though electric motors are highly efficient 10–20% of the power consumed is wasted, not being used for doing useful work. Most of the 10–20% of the power lost is transferred into heat dissipated by the copper wires used to make the coils of the motor. This heat adds to the heat of the environment and adds to the demand on power plants providing the power.

The demand on the power plant can lead to increased greenhouse gases, particularly if the power plant uses coal or gas as fuel.

A fuse ([link]) is a device that protects a circuit from currents that are too high. A fuse is basically a short piece of wire between two contacts. As we have seen, when a current is running through a conductor, the kinetic energy of the charge carriers is converted into thermal energy in the conductor. The piece of wire in the fuse is under tension and has a low melting point. The wire is designed to heat up and break at the rated current. The fuse is destroyed and must be replaced, but it protects the rest of the circuit. Fuses act quickly, but there is a small time delay while the wire heats up and breaks.

A fuse consists of a piece of wire between two contacts. When a current passes through the wire that is greater than the rated current, the wire melts, breaking the connection. Pictured is a “blown” fuse where the wire broke protecting a circuit (credit: modification of work by “Shardayyy”/Flickr).

Circuit breakers are also rated for a maximum current, and open to protect the circuit, but can be reset. Circuit breakers react much faster. The operation of circuit breakers is not within the scope of this chapter and will be discussed in later chapters. Another method of protecting equipment and people is the ground fault circuit interrupter (GFCI), which is common in bathrooms and kitchens. The GFCI outlets respond very quickly to changes in current. These outlets open when there is a change in magnetic field produced by current-carrying conductors, which is also beyond the scope of this chapter and is covered in a later chapter.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = \frac{dE}{dt}$, we see that

Equation:

$$E = \int P dt$$

is the energy used by a device using power P for a time interval t. If power is delivered at a constant rate, then then the energy can be found by $E = Pt$. For example, the more light bulbs burning, the greater P used; the longer they are on, the greater t is.

The energy unit on electric bills is the kilowatt-hour $(\text{kW} \cdot \text{h})$, consistent with the relationship $E = Pt$. It is easy to estimate the cost of operating electrical appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted into joules. You can prove to yourself that $1 \text{ kW} \cdot \text{h} = 3.6 \times 10^6 \text{ J}$.

The electrical energy (E) used can be reduced either by reducing the time of use or by reducing the power consumption of that appliance or fixture. This not only reduces the cost but also results in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home's use of energy goes to lighting, and the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFLs). (See [link](b).) Thus, a 60-W incandescent bulb can be replaced by a 15-W CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.)

The heat transfer from these CFLs is less, and they last up to 10 times longer than incandescent bulbs. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last five times longer than CFLs.

Example:

Calculating the Cost Effectiveness of LED Bulb

The typical replacement for a 100-W incandescent bulb is a 20-W LED bulb. The 20-W LED bulb can provide the same amount of light output as the 100-W incandescent light bulb. What is the cost savings for using the LED bulb in place of the incandescent bulb for one year, assuming $0.10 per kilowatt-hour is the average energy rate charged by the power company? Assume that the bulb is turned on for three hours a day.

Strategy

(a) Calculate the energy used during the year for each bulb, using $E = Pt$.

(b) Multiply the energy by the cost.

Solution

a. Calculate the power for each bulb.
   Equation:

$$\begin{aligned} E_{\text{Incandescent}} &= Pt = 100\text{ W}\left(\frac{1\text{ kW}}{1000\text{ W}}\right)\left(\frac{3\text{ h}}{\text{day}}\right)(365\text{ days}) = 109.5\text{ kW}\cdot\text{h} \\ E_{\text{LED}} &= Pt = 20\text{ W}\left(\frac{1\text{ kW}}{1000\text{ W}}\right)\left(\frac{3\text{ h}}{\text{day}}\right)(365\text{ days}) = 21.90\text{ kW}\cdot\text{h} \end{aligned}.$$

b. Calculate the cost for each.
   Equation:

$$\begin{aligned} \text{cost}_{\text{Incandescent}} &= 109.5\text{ kW-h}\left(\frac{\$0.10}{\text{kW}\cdot\text{h}}\right) = \$10.95 \\ \text{cost}_{\text{LED}} &= 21.90\text{ kW-h}\left(\frac{\$0.10}{\text{kW}\cdot\text{h}}\right) = \$2.19 \end{aligned}.$$

Significance

A LED bulb uses 80% less energy than the incandescent bulb, saving $8.76 over the incandescent bulb for one year. The LED bulb can cost $20.00 and the 100-W incandescent bulb can cost $0.75, which should be calculated into the computation. A typical lifespan of an incandescent bulb is 1200 hours and is 50,000 hours for the LED bulb. The incandescent bulb would last 1.08 years at 3 hours a day and the LED bulb would last 45.66 years. The initial cost of the LED bulb is high, but the cost to the home owner will be $0.69 for the incandescent bulbs versus $0.44 for the LED bulbs per year. (Note that the LED bulbs are coming down in price.) The cost savings per year is approximately $8.50, and that is just for one bulb.

Note:
Exercise:

Problem:

Check Your Understanding Is the efficiency of the various light bulbs the only consideration when comparing the various light bulbs?

Solution:

No, the efficiency is a very important consideration of the light bulbs, but there are many other considerations. As mentioned above, the cost of the bulbs and the life span of the bulbs are important considerations. For example, CFL bulbs contain mercury, a neurotoxin, and must be disposed of as hazardous waste. When replacing incandescent bulbs that are being controlled by a dimmer switch with LED, the dimmer switch may need to be replaced. The dimmer switches for LED lights are comparably priced to the incandescent light switches, but this is an initial cost which should be considered. The spectrum of light should also be considered, but there is a broad range of color temperatures available, so you should be able to find one that fits your needs. None of these considerations mentioned are meant to discourage the use of LED or CFL light bulbs, but they are considerations.

Changing light bulbs from incandescent bulbs to CFL or LED bulbs is a simple way to reduce energy consumption in homes and commercial sites. CFL bulbs operate with a much different mechanism than do incandescent lights. The mechanism is complex and beyond the scope of this chapter, but here is a very general description of the mechanism. CFL bulbs contain argon and mercury vapor housed within a spiral-shaped tube. The CFL bulbs use a “ballast” that increases the voltage used by the CFL bulb. The ballast produce an electrical current, which passes through the gas mixture and excites the gas molecules. The excited gas molecules produce ultraviolet (UV) light, which in turn stimulates the fluorescent coating on the inside of the tube. This coating fluoresces in the visible spectrum, emitting visible light. Traditional fluorescent tubes and CFL bulbs had a short time delay of up to a few seconds while the mixture was being “warmed up” and the molecules reached an excited state. It should be noted that these bulbs do contain mercury, which is poisonous, but if the bulb is broken, the mercury is never released. Even if the bulb is broken, the mercury tends to remain in the fluorescent coating. The amount is also quite small and the advantage of the energy saving may outweigh the disadvantage of using mercury.

The CFL light bulbs are being replaced with LED light bulbs, where LED stands for “light-emitting diode.” The diode was briefly discussed as a nonohmic device, made of semiconducting material, which essentially permits current flow in one direction. LEDs are a special type of diode made of semiconducting materials infused with impurities in combinations and concentrations that enable the extra energy from the movement of the electrons during electrical excitation to be converted into visible light. Semiconducting devices will be explained in greater detail in Condensed Matter Physics.

Commercial LEDs are quickly becoming the standard for commercial and residential lighting, replacing incandescent and CFL bulbs. They are designed for the visible spectrum and are constructed from gallium doped with arsenic and phosphorous atoms. The color emitted from an LED depends on the materials used in the semiconductor and the current. In the early years of LED development, small LEDs found on circuit boards were red, green, and yellow, but LED light bulbs can now be programmed to produce millions of colors of light as well as many different hues of white light.

Comparison of Incandescent, CFL, and LED Light Bulbs

The energy savings can be significant when replacing an incandescent light bulb or a CFL light bulb with an LED light. Light bulbs are rated by the amount of power that the bulb consumes, and the amount of light output is measured in lumens. The lumen (lm) is the SI -derived unit of luminous flux and is a measure of the total quantity of visible light emitted by a source. A 60-W incandescent light bulb can be replaced with a 13- to 15-W CFL bulb or a 6- to 8-W LED bulb, all three of which have a light output of approximately 800 lm. A table of light output for some commonly used light bulbs appears in [link].

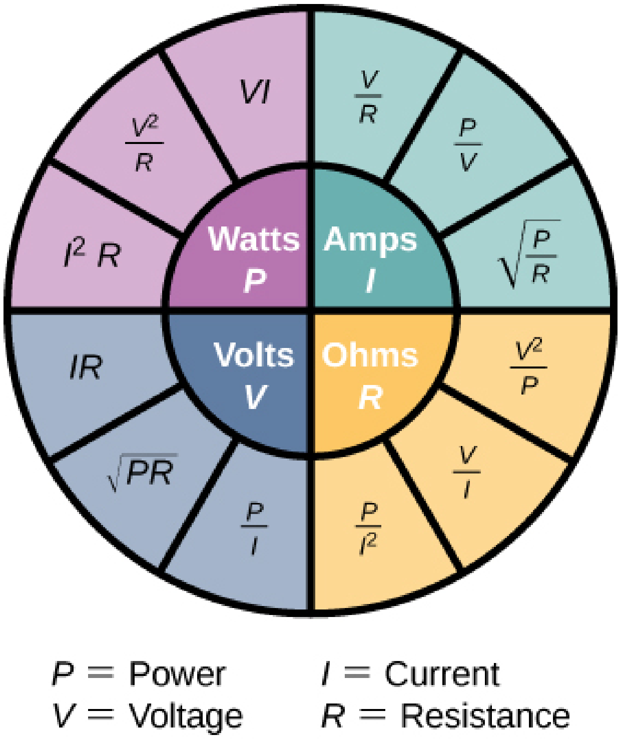
The life spans of the three types of bulbs are significantly different. An LED bulb has a life span of 50,000 hours, whereas the CFL has a lifespan of 8000 hours and the incandescent lasts a mere 1200 hours. The LED bulb is the most durable, easily withstanding rough treatment such as jarring and bumping. The incandescent light bulb has little tolerance to the same treatment since the filament and glass can easily break. The CFL bulb is also less durable than the LED bulb because of its glass construction. The amount of heat emitted is 3.4 btu/h for the 8-W LED bulb, 85 btu/h for the 60-W incandescent bulb, and 30 btu/h for the CFL bulb. As mentioned earlier, a major drawback of the CFL bulb is that it contains mercury, a neurotoxin, and must be disposed of as hazardous waste. From these data, it is easy to understand why the LED light bulb is quickly becoming the standard in lighting.

Light Output (lumens)	LED Light Bulbs (watts)	Incandescent Light Bulbs (watts)	CFL Light Bulbs (watts)
450	4−5	40	9−13
800	6−8	60	13−15
1100	9−13	75	18−25
1600	16−20	100	23−30
2600	25−28	150	30−55

Light Output of LED, Incandescent, and CFL Light Bulbs

Summary of Relationships

In this chapter, we have discussed relationships between voltages, current, resistance, and power. [link] shows a summary of the relationships between these measurable quantities for ohmic devices. (Recall that ohmic devices follow Ohm's law $V = IR$.) For example, if you need to calculate the power, use the pink section, which shows that $P = VI$, $P = \frac{V^2}{R}$, and $P = I^2R$.


VI
$\frac{V^2}{R}$
$I^2 R$
Watts
P
$\frac{V}{R}$
$\frac{P}{V}$
$\sqrt{\frac{P}{R}}$
Amps
I
IR
$\sqrt{PR}$
$\frac{P}{I}$
Volts
V
$\frac{V^2}{P}$
$\frac{V}{I}$
$\frac{P}{I^2}$
Ohms
R
P = Power
V = Voltage
I = Current
R = Resistance

This circle shows a summary of the equations for the relationships between power, current, voltage, and resistance.

Which equation you use depends on what values you are given, or you measure. For example if you are given the current and the resistance, use $P = I^2R$. Although all the possible combinations may seem overwhelming, don't forget that they all are combinations of just two equations, Ohm's law $(V = IR)$ and power $(P = IV)$.

Summary

- Electric power is the rate at which electric energy is supplied to a circuit or consumed by a load.
- Power dissipated by a resistor depends on the square of the current through the resistor and is equal to $P = I^2R = \frac{V^2}{R}$.
- The SI unit for electric power is the watt and the SI unit for electric energy is the joule. Another common unit for electric energy, used by power companies, is the kilowatt-hour (kW · h).
- The total energy used over a time interval can be found by $E = \int Pdt$.

Conceptual Questions

Exercise:

Problem:

Common household appliances are rated at 110 V, but power companies deliver voltage in the kilovolt range and then step the voltage down using transformers to 110 V to be used in homes. You will learn in later chapters that transformers consist of many turns of wire, which warm up as current flows through them, wasting some of the energy that is given off as heat. This sounds inefficient. Why do the power companies transport electric power using this method?

Solution:

Although the conductors have a low resistance, the lines from the power company can be kilometers long. Using a high voltage reduces the current that is required to supply the power demand and that reduces line losses.

Exercise:

Problem:

Your electric bill gives your consumption in units of kilowatt-hour (kW · h). Does this unit represent the amount of charge, current, voltage, power, or energy you buy?

Exercise:

Problem:

Resistors are commonly rated at $\frac{1}{8}\,\mathrm{W}$, $\frac{1}{4}\,\mathrm{W}$, $\frac{1}{2}\,\mathrm{W}$, 1 W and 2 W for use in electrical circuits. If a current of $I = 2.00\ \mathrm{A}$ is accidentally passed through a $R = 1.00\ \Omega$ resistor rated at 1 W, what would be the most probable outcome? Is there anything that can be done to prevent such an accident?

Solution:

The resistor would overheat, possibly to the point of causing the resistor to burn. Fuses are commonly added to circuits to prevent such accidents.

Exercise:

Problem:

An immersion heater is a small appliance used to heat a cup of water for tea by passing current through a resistor. If the voltage applied to the appliance is doubled, will the time required to heat the water change? By how much? Is this a good idea?

Problems

Exercise:

Problem:

A 20.00-V battery is used to supply current to a $10\text{-k}\,\Omega$ resistor. Assume the voltage drop across any wires used for connections is negligible. (a) What is the current through the resistor? (b) What is the power dissipated by the resistor? (c) What is the power input from the battery, assuming all the electrical power is dissipated by the resistor? (d) What happens to the energy dissipated by the resistor?

Solution:

a. $I = 2\ \mathrm{mA}$; b. $P = 0.04\ \mathrm{W}$; c. $P = 0.04\ \mathrm{W}$; d. It is converted into heat.

Exercise:

Problem:

What is the maximum voltage that can be applied to a $20\text{-k}\,\Omega$ resistor rated at $\frac{1}{4}\,\text{W}$?

Exercise:

Problem:

A heater is being designed that uses a coil of 14-gauge nichrome wire to generate 300 W using a voltage of $V = 110\,\text{V}$. How long should the engineer make the wire?

Solution:

$$\begin{array}{l} P = \frac{V^2}{R} \\ R = 40\,\Omega \end{array}, \begin{array}{l} A = 2.08\text{mm}^2 \\ \rho = 100 \times 10^{-8}\,\Omega\cdot\text{m} \\ R = \rho\frac{L}{A} \\ L = 83\,\text{m} \end{array}$$

Exercise:

Problem:

An alternative to CFL bulbs and incandescent bulbs are light-emitting diode (LED) bulbs. A 100-W incandescent bulb can be replaced by a 16-W LED bulb. Both produce 1600 lumens of light. Assuming the cost of electricity is $0.10 per kilowatt-hour, how much does it cost to run the bulb for one year if it runs for four hours a day?

Exercise:

Problem:

The power dissipated by a resistor with a resistance of $R = 100\,\Omega$ is $P = 2.0\,\text{W}$. What are the current through and the voltage drop across the resistor?

Solution:

$$I = 0.14\,\text{A}, \quad V = 14\,\text{V}$$

Exercise:

Problem:

Running late to catch a plane, a driver accidentally leaves the headlights on after parking the car in the airport parking lot. During takeoff, the driver realizes the mistake. Having just replaced the battery, the driver knows that the battery is a 12-V automobile battery, rated at 100 $\text{A}\cdot\text{h}$. The driver, knowing there is nothing that can be done, estimates how long the lights will shine, assuming there are two 12-V headlights, each rated at 40 W. What did the driver conclude?

Exercise:

Problem:

A physics student has a single-occupancy dorm room. The student has a small refrigerator that runs with a current of 3.00 A and a voltage of 110 V, a lamp that contains a 100-W bulb, an overhead light with a 60-W bulb, and various other small devices adding up to 3.00 W. (a) Assuming the power plant that supplies 110 V electricity to the dorm is 10 km away and the two aluminum transmission cables use 0-gauge wire with a diameter of 8.252 mm, estimate the percentage of the total power supplied by the power company that is lost in the transmission. (b) What would be the result is the power company delivered the electric power at 110 kV?

Solution:

a.
$$I \approx 3.00\text{ A} + \frac{100\text{ W}}{110\text{ V}} + \frac{60\text{ W}}{110\text{ V}} + \frac{3.00\text{ W}}{110\text{ V}} = 4.48\text{ A}$$
$$P = 493\text{ W}$$
$$R = 9.91\ \Omega\,,$$
$$P_{\text{loss}} = 200.\text{ W}$$
$$\%\text{loss} = 40\%$$

b.
$$P = 493\text{ W}$$
$$I = 0.0045\text{ A}$$
$$R = 9.91\ \Omega$$
$$P_{\text{loss}} = 201\mu\text{ W}$$
$$\%\text{loss} = 0.00004\%$$

Exercise:

Problem:

A 0.50-W, 220- Ω resistor carries the maximum current possible without damaging the resistor. If the current were reduced to half the value, what would be the power consumed?

Glossary

electrical power	time rate of change of energy in an electric circuit

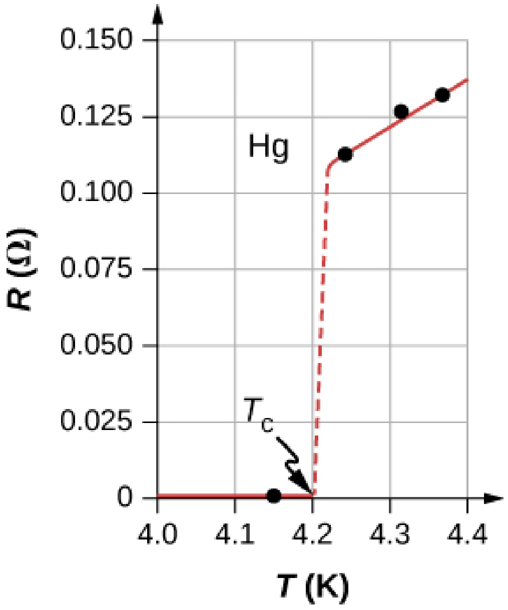
Superconductors
By the end of this section, you will be able to:

- Describe the phenomenon of superconductivity
- List applications of superconductivity

Touch the power supply of your laptop computer or some other device. It probably feels slightly warm. That heat is an unwanted byproduct of the process of converting household electric power into a current that can be used by your device. Although electric power is reasonably efficient, other losses are associated with it. As discussed in the section on power and energy, transmission of electric power produces I^2R line losses. These line losses exist whether the power is generated from conventional power plants (using coal, oil, or gas), nuclear plants, solar plants, hydroelectric plants, or wind farms. These losses can be reduced, but not eliminated, by transmitting using a higher voltage. It would be wonderful if these line losses could be eliminated, but that would require transmission lines that have zero resistance. In a world that has a global interest in not wasting energy, the reduction or elimination of this unwanted thermal energy would be a significant achievement. Is this possible?

The Resistance of Mercury

In 1911, Heike Kamerlingh Onnes of Leiden University, a Dutch physicist, was looking at the temperature dependence of the resistance of the element mercury. He cooled the sample of mercury and noticed the familiar behavior of a linear dependence of resistance on temperature; as the temperature decreased, the resistance decreased. Kamerlingh Onnes continued to cool the sample of mercury, using liquid helium. As the temperature approached $4.2\ \text{K}\ (-269.2\ ^\circ\text{C})$, the resistance abruptly went to zero ([link]). This temperature is known as the **critical temperature** T_c for mercury. The sample of mercury entered into a phase where the resistance was absolutely zero. This phenomenon is known as **superconductivity**. (*Note:* If you connect the leads of a three-digit ohmmeter across a conductor, the reading commonly shows up as $0.00\ \Omega$. The resistance of the conductor is not actually zero, it is less than $0.01\ \Omega$.) There are various methods to measure very small resistances, such as the four-point method, but an ohmmeter is not an acceptable method to use for testing resistance in superconductivity.



The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to the temperature of about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

Other Superconducting Materials

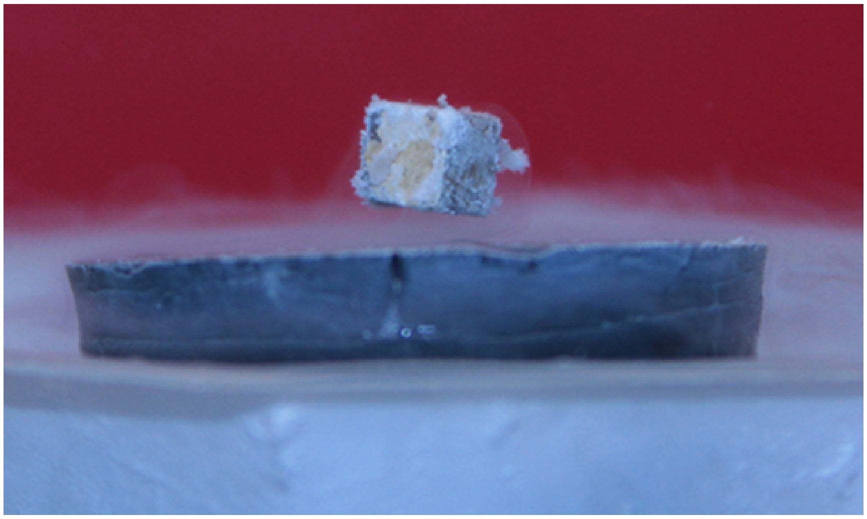
As research continued, several other materials were found to enter a superconducting phase, when the temperature reached near absolute zero. In 1941, an alloy of niobium-nitride was found that could become superconducting at $T_{\text{c}} = 16\,\text{K}\,(-257\,^{\circ}\text{C})$ and in 1953, vanadium-silicon was found to become superconductive at $T_{\text{c}} = 17.5\,\text{K}(-255.7\,^{\circ}\text{C})$. The temperatures for the transition into superconductivity were slowly creeping higher. Strangely, many materials that make good conductors, such as copper, silver, and gold, do not exhibit superconductivity. Imagine the energy savings if transmission lines for electric power-generating stations could be made to be superconducting at temperatures near room temperature! A resistance of zero ohms means no I^2R losses and a great boost to reducing energy consumption. The problem is that $T_{\text{c}} = 17.5\,\text{K}$ is still very cold and in the range of liquid helium temperatures. At this temperature, it is not cost effective to transmit electrical energy because of the cooling requirements.

A large jump was seen in 1986, when a team of researchers, headed by Dr. Ching Wu Chu of Houston University, fabricated a brittle, ceramic compound with a transition temperature of $T_{\text{c}} = 92\,\text{K}\,(-181\,^{\circ}\text{C})$. The ceramic material, composed of yttrium barium copper oxide (YBCO), was an insulator at room temperature. Although this temperature still seems quite cold, it is near the boiling point of liquid nitrogen, a liquid commonly used in refrigeration. You may have noticed refrigerated trucks traveling down the highway labeled as "Liquid Nitrogen Cooled."

YBCO ceramic is a material that could be useful for transmitting electrical energy because the cost saving of reducing the I^2R losses are larger than the cost of cooling the superconducting cable, making it financially feasible. There were and are many engineering problems to overcome. For example, unlike traditional electrical cables, which are flexible and have a decent tensile strength, ceramics are brittle and would break rather than stretch under pressure. Processes that are rather simple with traditional cables, such as making connections, become difficult when working with ceramics. The problems are difficult and complex, and material scientists and engineers are coming up with innovative solutions.

An interesting consequence of the resistance going to zero is that once a current is established in a superconductor, it persists without an applied voltage source. Current loops in a superconductor have been set up and the current loops have been observed to persist for years without decaying.

Zero resistance is not the only interesting phenomenon that occurs as the materials reach their transition temperatures. A second effect is the exclusion of magnetic fields. This is known as the **Meissner effect** ([link]). A light, permanent magnet placed over a superconducting sample will levitate in a stable position above the superconductor. High-speed trains have been developed that levitate on strong superconducting magnets, eliminating the friction normally experienced between the train and the tracks. In Japan, the Yamanashi Maglev test line opened on April 3, 1997. In April 2015, the MLX01 test vehicle attained a speed of 374 mph (603 km/h).



A small, strong magnet levitates over a superconductor cooled to liquid nitrogen temperature. The magnet levitates because the superconductor excludes magnetic fields. (credit: Joseph J. Trout)

[link] shows a select list of elements, compounds, and high-temperature superconductors, along with the critical temperatures for which they become superconducting. Each section is sorted from the highest critical temperature to the lowest. Also listed is the critical magnetic field for some of the materials. This is the strength of the magnetic field that destroys superconductivity. Finally, the type of the superconductor is listed.

There are two types of superconductors. There are 30 pure metals that exhibit zero resistivity below their critical temperature and exhibit the Meissner effect, the property of excluding magnetic fields from the interior of the superconductor while the superconductor is at a temperature below the critical temperature. These metals are called Type I superconductors. The superconductivity exists only below their critical temperatures and below a critical magnetic field strength. Type I superconductors are well described by the BCS theory (described next). Type I superconductors have limited practical applications because the strength of the critical magnetic field needed to destroy the superconductivity is quite low.

Type II superconductors are found to have much higher critical magnetic fields and therefore can carry much higher current densities while remaining in the superconducting state. A collection of various ceramics containing barium-copper-oxide have much higher critical temperatures for the transition into a superconducting state. Superconducting materials that belong to this subcategory of the Type II superconductors are often categorized as high-temperature superconductors.

Introduction to BCS Theory

Type I superconductors, along with some Type II superconductors can be modeled using the BCS theory, proposed by John Bardeen, Leon Cooper, and Robert Schrieffer. Although the theory is beyond the scope of this chapter, a short summary of the theory is provided here. (More detail is provided in Condensed Matter Physics.) The theory considers pairs of electrons and how they are coupled together through lattice-vibration interactions. Through the interactions with the crystalline lattice, electrons near the Fermi energy level feel a small attractive force and form pairs (Cooper pairs), and the coupling is known as a phonon interaction. Single electrons are fermions, which are particles that obey the Pauli exclusion principle. The Pauli exclusion principle in quantum mechanics states that two identical fermions (particles with half-integer spin) cannot occupy the same quantum state simultaneously. Each electron has four quantum numbers (n, l, m_l, m_s). The principal quantum number (n) describes the energy of the electron, the orbital angular momentum quantum number (l) indicates the most probable distance from the nucleus, the magnetic quantum number (m_l) describes the energy levels in the subshell, and the electron spin quantum number (m_s) describes the orientation of the spin of the electron, either up or down. As the material enters a superconducting state, pairs of electrons act more like bosons, which can condense into the same energy level and need not obey the Pauli exclusion principle. The electron pairs have a slightly lower energy and leave an energy gap above them on the order of 0.001 eV. This energy gap inhibits collision interactions that lead to ordinary resistivity. When the material is below the critical temperature, the thermal energy is less than the band gap and the material exhibits zero resistivity.

Material	**Symbol or Formula**	**Critical Temperature T_c (K)**	**Critical Magnetic Field H_c (T)**	**Type**
Elements				

Material	**Symbol or Formula**	**Critical Temperature T_c (K)**	**Critical Magnetic Field H_c (T)**	**Type**
Lead	Pb	7.19	0.08	I
Lanthanum	La	(α) 4.90 − (β) 6.30		I
Tantalum	Ta	4.48	0.09	I
Mercury	Hg	(α) 4.15 − (β) 3.95	0.04	I
Tin	Sn	3.72	0.03	I
Indium	In	3.40	0.03	I
Thallium	Tl	2.39	0.03	I
Rhenium	Re	2.40	0.03	I
Thorium	Th	1.37	0.013	I
Protactinium	Pa	1.40		I
Aluminum	Al	1.20	0.01	I
Gallium	Ga	1.10	0.005	I
Zinc	Zn	0.86	0.014	I
Titanium	Ti	0.39	0.01	I
Uranium	U	(α) 0.68 − (β) 1.80		I
Cadmium	Cd	11.4	4.00	I
Compounds				

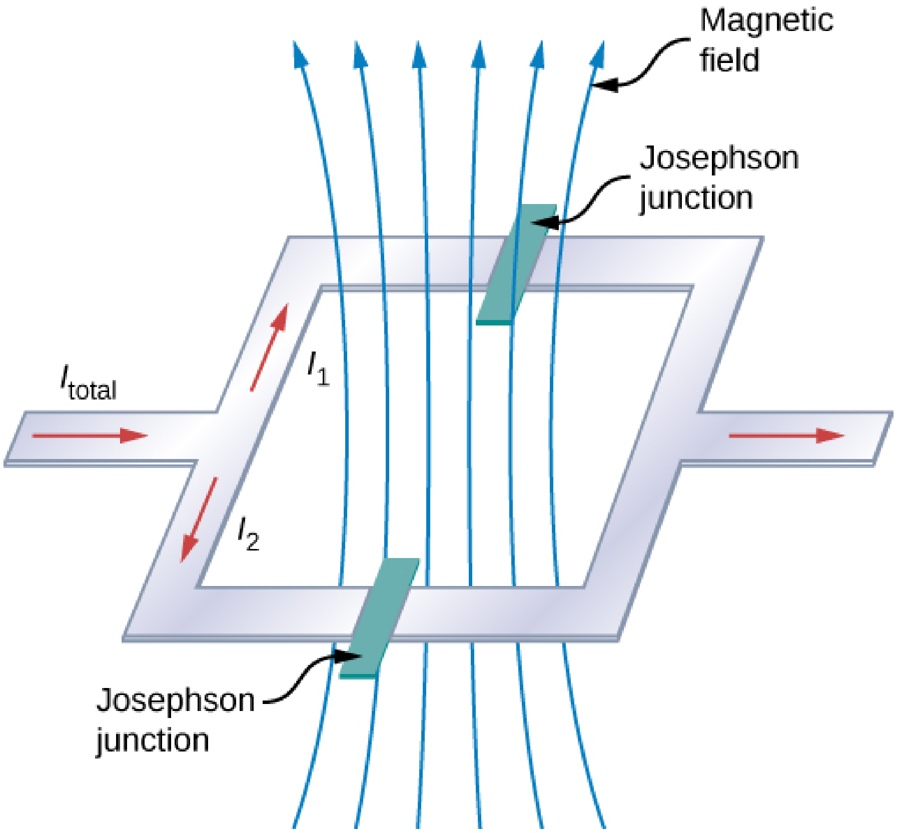
Material	**Symbol or Formula**	**Critical Temperature T_c (K)**	**Critical Magnetic Field H_c (T)**	**Type**
Niobium-germanium	Nb_3Ge	23.20	37.00	II
Niobium-tin	Nb_3Sn	18.30	30.00	II
Niobium-nitrite	NbN	16.00		II
Niobium-titanium	NbTi	10.00	15.00	II
High-Temperature Oxides				
	$HgBa_2CaCu_2O_8$	134.00		II
	$Tl_2Ba_2Ca_2Cu_3O_{10}$	125.00		II
	$YBa_2Cu_3O_7$	92.00	120.00	II

Superconductor Critical Temperatures

Applications of Superconductors

Superconductors can be used to make superconducting magnets. These magnets are 10 times stronger than the strongest electromagnets. These magnets are currently in use in magnetic resonance imaging (MRI), which produces high-quality images of the body interior without dangerous radiation.

Another interesting application of superconductivity is the **SQUID** (superconducting quantum interference device). A SQUID is a very sensitive magnetometer used to measure extremely subtle magnetic fields. The operation of the SQUID is based on superconducting loops containing Josephson junctions. A **Josephson junction** is the result of a theoretical prediction made by B. D. Josephson in an article published in 1962. In the article, Josephson described how a supercurrent can flow between two pieces of superconductor separated by a thin layer of insulator. This phenomenon is now called the Josephson effect. The SQUID consists of a superconducting current loop containing two Josephson junctions, as shown in [link]. When the loop is placed in even a very weak magnetic field, there is an interference effect that depends on the strength of the magnetic field.



The SQUID (superconducting quantum interference device) uses a superconducting current loop and two Josephson junctions to detect magnetic fields as low as 10^{-14} T (Earth’s magnet field is on the order of 0.3×10^{-5} T).

Superconductivity is a fascinating and useful phenomenon. At critical temperatures near the boiling point of liquid nitrogen, superconductivity has special applications in MRIs, particle accelerators, and high-speed trains. Will we reach a state where we can have materials enter the superconducting phase at near room temperatures? It seems a long way off, but if scientists in 1911 were asked if we would reach liquid-nitrogen temperatures with a ceramic, they might have thought it implausible.

Summary

- Superconductivity is a phenomenon that occurs in some materials when cooled to very low critical temperatures, resulting in a resistance of exactly zero and the expulsion of all magnetic fields.
- Materials that are normally good conductors (such as copper, gold, and silver) do not experience superconductivity.
- Superconductivity was first observed in mercury by Heike Kamerlingh Onnes in 1911. In 1986, Dr. Ching Wu Chu of Houston University fabricated a brittle, ceramic compound with a critical temperature close to the temperature of liquid nitrogen.
- Superconductivity can be used in the manufacture of superconducting magnets for use in MRIs and high-speed, levitated trains.

Key Equations

Average electrical current	$I_{\text{ave}} = \frac{\Delta Q}{\Delta t}$
Definition of an ampere	$1\ \text{A} = 1\ \text{C/s}$
Electrical current	$I = \frac{dQ}{dt}$
Drift velocity	$v_d = \frac{I}{nqA}$
Current density	$I = \iint\limits_{\text{area}} \vec{\mathbf{J}} \cdot d\vec{\mathbf{A}}$
Resistivity	$\rho = \frac{E}{J}$
Common expression of Ohm’s law	$V = IR$
Resistivity as a function of temperature	$\rho = \rho_0 \left[1 + \alpha \left(T - T_0\right)\right]$
Definition of resistance	$R \equiv \frac{V}{I}$

Resistance of a cylinder of material	$R=\rho\frac{L}{A}$
Temperature dependence of resistance	$R=R_0\left(1+\alpha\Delta T\right)$
Electric power	$P=IV$
Power dissipated by a resistor	$P=I^2R=\frac{V^2}{R}$

Conceptual Questions

Exercise:

Problem:

What requirement for superconductivity makes current superconducting devices expensive to operate?

Solution:

Very low temperatures necessitate refrigeration. Some materials require liquid nitrogen to cool them below their critical temperatures. Other materials may need liquid helium, which is even more costly.

Exercise:

Problem:

Name two applications for superconductivity listed in this section and explain how superconductivity is used in the application. Can you think of a use for superconductivity that is not listed?

Problems

Exercise:

Problem:

Consider a power plant is located 60 km away from a residential area uses 0-gauge $\left(A=42.40\,\text{mm}^2\right)$ wire of copper to transmit power at a current of $I=100.00\ \text{A}$. How much more power is dissipated in the copper wires than it would be in superconducting wires?

Solution:

$$R_{\text{copper}} = 23.77\,\Omega$$
$$P = 2.377\ \times\ 10^5\text{W}$$

Exercise:

Problem:

A wire is drawn through a die, stretching it to four times its original length. By what factor does its resistance increase?

Exercise:

Problem:

Digital medical thermometers determine temperature by measuring the resistance of a semiconductor device called a thermistor (which has $\alpha = -0.06/^\circ\text{C}$) when it is at the same temperature as the patient. What is a patient’s temperature if the thermistor’s resistance at that temperature is 82.0% of its value at $37\ ^\circ\text{C}$ (normal body temperature)?

Solution:

$$R = R_0\left(1 + \alpha\left(T - T_0\right)\right)$$
$$0.82R_0 = R_0\left(1 + \alpha\left(T - T_0\right)\right),\quad 0.82 = 1 – 0.06\left(T – 37\ ^\circ\text{C}\right),\quad T = 40\ ^\circ\text{C}$$

Exercise:

Problem:

Electrical power generators are sometimes “load tested” by passing current through a large vat of water. A similar method can be used to test the heat output of a resistor. A $R = 30\ \Omega$ resistor is connected to a 9.0-V battery and the resistor leads are waterproofed and the resistor is placed in 1.0 kg of room temperature water $(T = 20\ ^\circ\text{C})$. Current runs through the resistor for 20 minutes. Assuming all the electrical energy dissipated by the resistor is converted to heat, what is the final temperature of the water?

Exercise:

Problem:

A 12-gauge gold wire has a length of 1 meter. (a) What would be the length of a silver 12-gauge wire with the same resistance? (b) What are their respective resistances at the temperature of boiling water?

Solution:

a. $R_{\text{Au}} = R_{\text{Ag}}, \quad \rho_{\text{Au}}\frac{L_{\text{Au}}}{A_{\text{Au}}} = \rho_{\text{Ag}}\frac{L_{\text{Ag}}}{A_{\text{Ag}}}, \quad L_{\text{Ag}} = 1.53\,\text{m};$
b. $R_{\text{Au,20}\,^\circ\text{C}} = 0.0074\,\Omega\,, \quad R_{\text{Au,100}\,^\circ\text{C}} = 0.0094\,\Omega\,, \quad R_{\text{Ag,100}\,^\circ\text{C}} = 0.0096\,\Omega$

Exercise:

Problem:

What is the change in temperature required to decrease the resistance for a carbon resistor by 10%?

Additional Problems

Exercise:

Problem:

A coaxial cable consists of an inner conductor with radius $r_{\text{i}} = 0.25\,\text{cm}$ and an outer radius of $r_{\text{o}} = 0.5\,\text{cm}$ and has a length of 10 meters. Plastic, with a resistivity of $\rho = 2.00 \times 10^{13}\,\Omega \cdot \text{m}$, separates the two conductors. What is the resistance of the cable?

Solution:

$$\begin{aligned} dR &= \frac{\rho}{2\pi rL}dr \\ R &= \frac{\rho}{2\pi L}\ln\frac{r_{\text{o}}}{r_{\text{i}}} \\ R &= 2.21 \times 10^{11}\,\Omega \end{aligned}$$

Exercise:

Problem:

A 10.00-meter long wire cable that is made of copper has a resistance of 0.051 ohms. (a) What is the weight if the wire was made of copper? (b) What is the weight of a 10.00-meter-long wire of the same gauge made of aluminum? (c)What is the resistance of the aluminum wire? The density of copper is $8960\,\text{kg/m}^3$ and the density of aluminum is $2760\,\text{kg/m}^3$.

Exercise:

Problem:

A nichrome rod that is 3.00 mm long with a cross-sectional area of $1.00\ \mathrm{mm}^2$ is used for a digital thermometer. (a) What is the resistance at room temperature? (b) What is the resistance at body temperature?

Solution:

a.
$R_0 = 0.003\ \Omega$; b.

$$\begin{aligned} T_\mathrm{c} &= 37.0\ ^\circ\mathrm{C} \\ R &= 0.00302\ \Omega \end{aligned}$$

Exercise:

Problem:

The temperature in Philadelphia, PA can vary between $68.00\ ^\circ\mathrm{F}$ and $100.00\ ^\circ\mathrm{F}$ in one summer day. By what percentage will an aluminum wire's resistance change during the day?

Exercise:

Problem:

When 100.0 V is applied across a 5-gauge (diameter 4.621 mm) wire that is 10 m long, the magnitude of the current density is $2.0\ \times\ 10^8 \mathrm{A/m}^2$. What is the resistivity of the wire?

Solution:

$\rho = 5.00\ \times\ 10^{-8}\ \Omega\ \cdot \mathrm{m}$

Exercise:

Problem:

A wire with a resistance of $5.0\ \Omega$ is drawn out through a die so that its new length is twice times its original length. Find the resistance of the longer wire. You may assume that the resistivity and density of the material are unchanged.

Exercise:

Problem:

What is the resistivity of a wire of 5-gauge wire $(A = 16.8\ \times\ 10^{-6} \mathrm{m}^2)$, 5.00 m length, and $5.10\ \mathrm{m}\ \Omega$ resistance?

Solution:

$\rho = 1.71 \times 10^{-8}\,\Omega \cdot \text{m}$

Exercise:

Problem:

Coils are often used in electrical and electronic circuits. Consider a coil which is formed by winding 1000 turns of insulated 20-gauge copper wire (area $0.52\,\text{mm}^2$) in a single layer on a cylindrical non-conducting core of radius 2.0 mm. What is the resistance of the coil? Neglect the thickness of the insulation.

Exercise:

Problem:

Currents of approximately 0.06 A can be potentially fatal. Currents in that range can make the heart fibrillate (beat in an uncontrolled manner). The resistance of a dry human body can be approximately $100\,\text{k}\,\Omega$. (a) What voltage can cause 0.06 A through a dry human body? (b) When a human body is wet, the resistance can fall to $100\,\Omega$. What voltage can cause harm to a wet body?

Solution:

a. $V = 6000\text{ V}$; b. $V = 6\text{ V}$

Exercise:

Problem:

A 20.00-ohm, 5.00-watt resistor is placed in series with a power supply. (a) What is the maximum voltage that can be applied to the resistor without harming the resistor? (b) What would be the current through the resistor?

Exercise:

Problem:

A battery with an emf of 24.00 V delivers a constant current of 2.00 mA to an appliance. How much work does the battery do in three minutes?

Solution:

$P = \frac{W}{t}, \quad W = 8.64\,\text{J}$

Exercise:

Problem:

A 12.00-V battery has an internal resistance of a tenth of an ohm. (a) What is the current if the battery terminals are momentarily shorted together? (b) What is the terminal voltage if the battery delivers 0.25 amps to a circuit?

Challenge Problems

Exercise:

Problem:

A 10-gauge copper wire has a cross-sectional area $A = 5.26\ \text{mm}^2$ and carries a current of $I = 5.00\ \text{A}$. The density of copper is $\rho = 8.95\ \text{g/cm}^3$. One mole of copper atoms $\left(6.02\ \times\ 10^{23}\text{atoms}\right)$ has a mass of approximately 63.50 g. What is the magnitude of the drift velocity of the electrons, assuming that each copper atom contributes one free electron to the current?

Solution:

$$V = 7.09\ \text{cm}^3$$
$$n = 8.49\ \times\ 10^{28}\,\frac{\text{electrons}}{\text{m}^3}$$
$$v_\text{d} = 7.00\ \times\ 10^{-5}\,\frac{\text{m}}{\text{s}}$$

Exercise:

Problem:

The current through a 12-gauge wire is given as $I\left(t\right) = \left(5.00\ \text{A}\right)\sin\ \left(2\pi 60\ \text{Hz}\ t\right)$. What is the current density at time 15.00 ms?

Exercise:

Problem:

A particle accelerator produces a beam with a radius of 1.25 mm with a current of 2.00 mA. Each proton has a kinetic energy of 10.00 MeV. (a) What is the velocity of the protons? (b) What is the number (n) of protons per unit volume? (b) How many electrons pass a cross sectional area each second?

Solution:

a. 4.38 x 10^7 m/s b. $v = 5.81\ \times\ 10^{13}\ \frac{\text{protons}}{\text{m}^3}$ c. $1.25\ \frac{\text{electrons}}{\text{m}^3}$

Exercise:

Problem:

In this chapter, most examples and problems involved direct current (DC). DC circuits have the current flowing in one direction, from positive to negative. When the current was changing, it was changed linearly from $I = -I_{\text{max}}$ to $I = +I_{\text{max}}$ and the voltage changed linearly from $V = -V_{\text{max}}$ to $V = +V_{\text{max}}$, where $V_{\text{max}} = I_{\text{max}}R$. Suppose a voltage source is placed in series with a resistor of $R = 10\ \Omega$ that supplied a current that alternated as a sine wave, for example, $I\left(t\right) = \left(3.00\ \text{A}\right)\sin\ \left(\frac{2\pi}{4.00\ \text{s}}t\right)$. (a) What would a graph of the voltage drop across the resistor *V*(*t*) versus time look like? (b) What would a plot of *V*(*t*) versus *I*(*t*) for one period look like? (*Hint*: If you are not sure, try plotting *V*(*t*) versus *I*(*t*) using a spreadsheet.)

Exercise:

Problem:

A current of $I = 25A$ is drawn from a 100-V battery for 30 seconds. By how much is the chemical energy reduced?

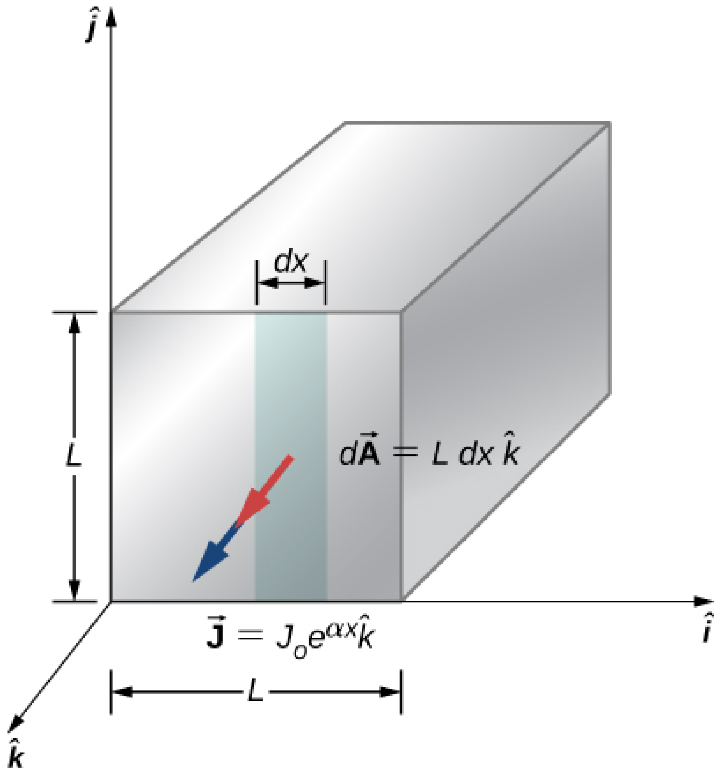
Solution:

$E = 75\ \text{kJ}$

Exercise:

Problem:

Consider a square rod of material with sides of length $L = 3.00\ \text{cm}$ with a current density of $\vec{\mathbf{J}} = J_0 e^{\alpha x}\widehat{k} = \left(0.35\frac{\text{A}}{\text{m}^2}\right)e^{\left(2.1\ \times\ 10^{-3}\text{m}^{-1}\right)x}\widehat{\mathbf{k}}$ as shown below. Find the current that passes through the face of the rod.


$\hat{j}$
dx
L
$d\vec{\mathbf{A}} = L\,dx\,\hat{k}$
$\vec{\mathbf{J}} = J_o e^{\alpha x}\hat{k}$
$\hat{i}$
L
$\hat{k}$

Exercise:

Problem:

A resistor of an unknown resistance is placed in an insulated container filled with 0.75 kg of water. A voltage source is connected in series with the resistor and a current of 1.2 amps flows through the resistor for 10 minutes. During this time, the temperature of the water is measured and the temperature change during this time is $\Delta T = 10.00\ ^\circ\text{C}$. (a) What is the resistance of the resistor? (b) What is the voltage supplied by the power supply?

Solution:

a. $P = 52\ \text{W}$; b. $V = 43.54\ \text{V}$
$R = 36\ \Omega$

Exercise:

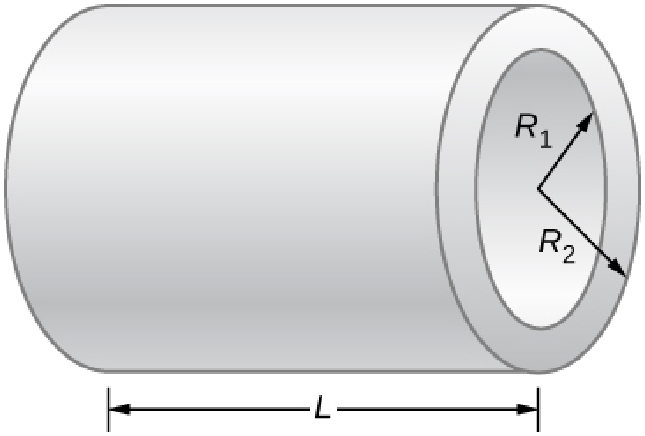
Problem:

The charge that flows through a point in a wire as a function of time is modeled as $q\left(t\right)=q_0e^{-t/T}=10.0\,\text{C}e^{-t/5\,\text{s}}$. (a) What is the initial current through the wire at time $t=0.00\,\text{s}$? (b) Find the current at time $t=\frac{1}{2}T$. (c) At what time t will the current be reduced by one-half $I=\frac{1}{2}I_0$?

Exercise:

Problem:

Consider a resistor made from a hollow cylinder of carbon as shown below. The inner radius of the cylinder is $R_i=0.20\,\text{mm}$ and the outer radius is $R_0=0.30\,\text{mm}$. The length of the resistor is $L=0.90\,\text{mm}$. The resistivity of the carbon is $\rho=3.5\,\times\,10^{-5}\,\Omega\,\cdot\text{m}$. (a) Prove that the resistance perpendicular from the axis is $R=\frac{\rho}{2\pi L}\text{ln}\left(\frac{R_0}{R_i}\right)$. (b) What is the resistance?



Solution:

a. $R=\frac{\rho}{2\pi L}\text{ln}\left(\frac{R_0}{R_i}\right)$; b. $R=2.5\,\text{m}\,\Omega$

Exercise:

Problem:

What is the current through a cylindrical wire of radius $R=0.1\,\text{mm}$ if the current density is $J=\frac{J_0}{R}r$, where $J_0=32000\frac{\text{A}}{\text{m}^2}$?

Exercise:

Problem:

A student uses a 100.00-W, 115.00-V radiant heater to heat the student's dorm room, during the hours between sunset and sunrise, 6:00 p.m. to 7:00 a.m. (a) What current does the heater operate at? (b) How many electrons move through the heater? (c) What is the resistance of the heater? (d) How much heat was added to the dorm room?

Solution:

(a) 0.870 A
(b) #electrons = 2.54×10^{23} electrons
(c) 132 ohms
(d) $q = 4.68 \times 10^{6}$ J

Exercise:

Problem:

A 12-V car battery is used to power a 20.00-W, 12.00-V lamp during the physics club camping trip/star party. The cable to the lamp is 2.00 meters long, 14-gauge copper wire with a charge density of $n = 9.50 \times 10^{28} \mathrm{m}^{-3}$. (a) What is the current draw by the lamp? (b) How long would it take an electron to get from the battery to the lamp?

Exercise:

Problem:

A physics student uses a 115.00-V immersion heater to heat 400.00 grams (almost two cups) of water for herbal tea. During the two minutes it takes the water to heat, the physics student becomes bored and decides to figure out the resistance of the heater. The student starts with the assumption that the water is initially at the temperature of the room $T_i = 25.00\ ^{\circ}\mathrm{C}$ and reaches $T_f = 100.00\ ^{\circ}\mathrm{C}$. The specific heat of the water is $c = 4180 \frac{\mathrm{J}}{\mathrm{kg} \bullet \mathrm{K}}$. What is the resistance of the heater?

Solution:

$$P = 1045\ \mathrm{W}, \quad P = \frac{V^2}{R}, \quad R = 12.27\,\Omega$$

Glossary

critical temperature
temperature at which a material reaches superconductivity

Josephson junction
junction of two pieces of superconducting material separated by a thin layer of insulating material, which can carry a supercurrent

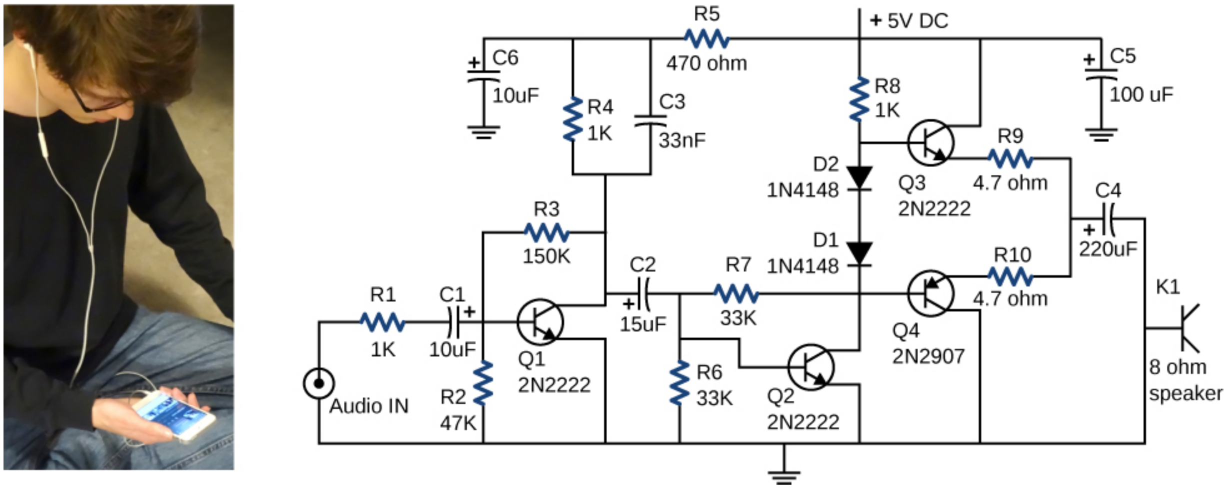
Meissner effect
phenomenon that occurs in a superconducting material where all magnetic fields are expelled

SQUID
(Superconducting Quantum Interference Device) device that is a very sensitive magnetometer, used to measure extremely subtle magnetic fields

superconductivity
phenomenon that occurs in some materials where the resistance goes to exactly zero and all magnetic fields are expelled, which occurs dramatically at some low critical temperature (T_C)

Introduction class="introduction"

This circuit shown is used to amplify small signals and power the earbud speakers attached to a cellular phone. This circuit’s components include resistors, capacitors, and diodes, all of which have been covered in previous chapters, as well as transistors, which are semi-conducting devices covered in Condensed Matter Physics. Circuits using similar components are found in all types of equipment and appliances you encounter in everyday life, such as alarm clocks, televisions, computers, and refrigerators. (credit left: modification of work by Jane Whitney)


R5
470 ohm
+ 5V DC
C6
10uF
R4
1K
C3
33nF
R8
1K
C5
100 uF
D2
1N4148
Q3
2N2222
R9
4.7 ohm
C4
220uF
R3
150K
D1
1N4148
R10
4.7 ohm
C2
15uF
R7
33K
K1
R1
1K
C1
10uF
Q1
2N2222
Q4
2N2907
8 ohm
speaker
Audio IN
R2
47K
R6
33K
Q2
2N2222

In the preceding few chapters, we discussed electric components, including capacitors, resistors, and diodes. In this chapter, we use these electric components in circuits. A circuit is a collection of electrical components connected to accomplish a specific task. [link] shows an amplifier circuit, which takes a small-amplitude signal and amplifies it to power the speakers in earbuds. Although the circuit looks complex, it actually consists of a set of series, parallel, and series-parallel circuits. The second section of this chapter covers the analysis of series and parallel circuits that consist of resistors. Later in this chapter, we introduce the basic equations and techniques to analyze any circuit, including those that are not reducible through simplifying parallel and series elements. But first, we need to understand how to power a circuit.

Electromotive Force

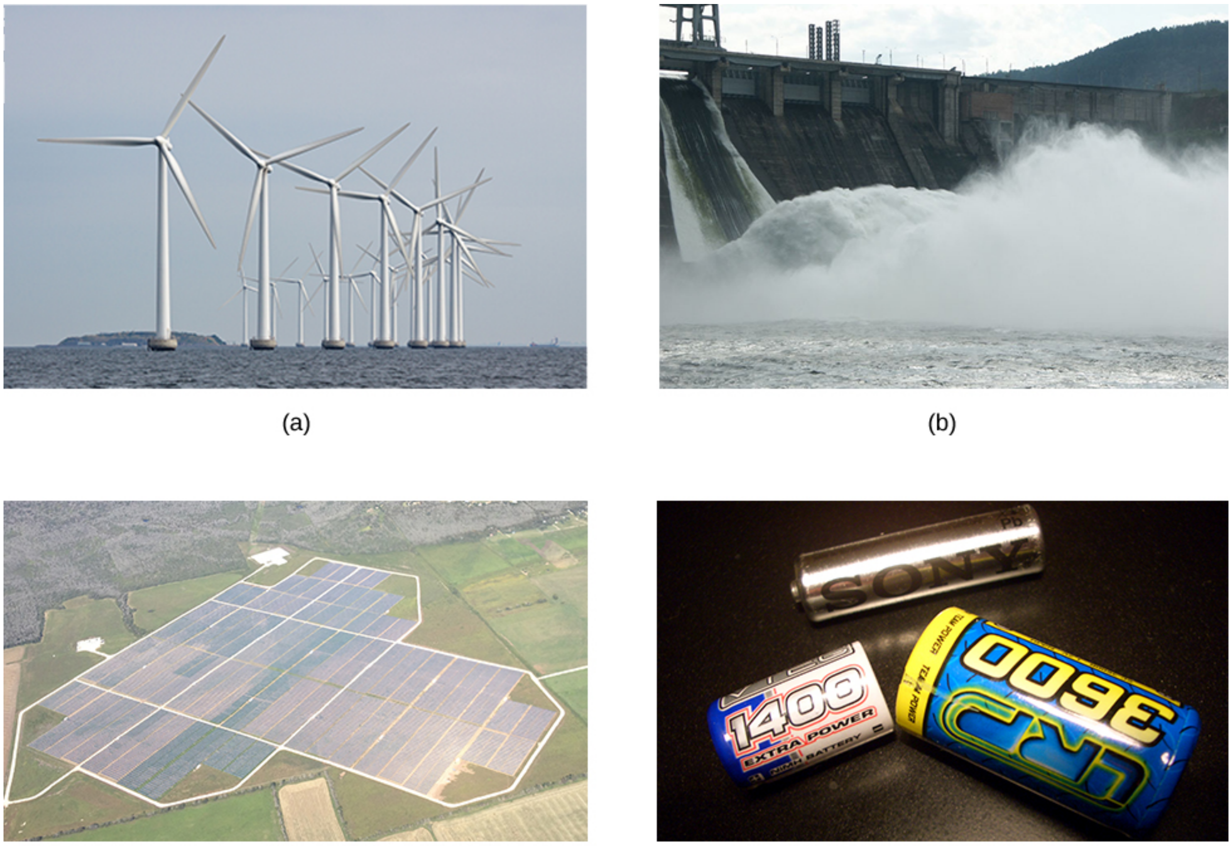
By the end of the section, you will be able to:

- Describe the electromotive force (emf) and the internal resistance of a battery
- Explain the basic operation of a battery

If you forget to turn off your car lights, they slowly dim as the battery runs down. Why don’t they suddenly blink off when the battery’s energy is gone? Their gradual dimming implies that the battery output voltage decreases as the battery is depleted. The reason for the decrease in output voltage for depleted batteries is that all voltage sources have two fundamental parts—a source of electrical energy and an internal resistance. In this section, we examine the energy source and the internal resistance.

Introduction to Electromotive Force

Voltage has many sources, a few of which are shown in [link]. All such devices create a **potential difference** and can supply current if connected to a circuit. A special type of potential difference is known as **electromotive force (emf)**. The emf is not a force at all, but the term ‘electromotive force’ is used for historical reasons. It was coined by Alessandro Volta in the 1800s, when he invented the first battery, also known as the voltaic pile. Because the electromotive force is not a force, it is common to refer to these sources simply as sources of emf (pronounced as the letters “ee-em-eff”), instead of sources of electromotive force.

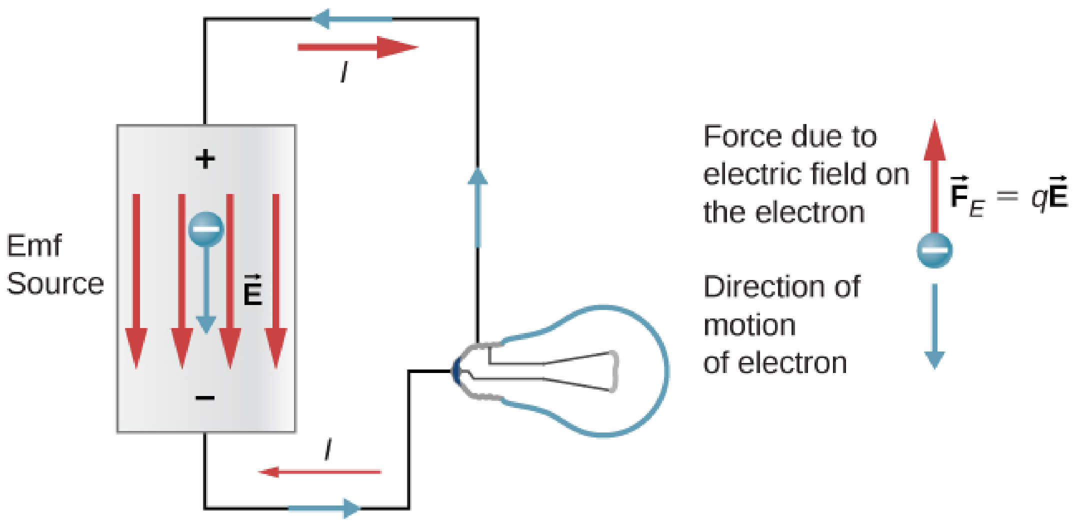


(a) (b)

(c) (d)

A variety of voltage sources. (a) The Brazos Wind Farm in Fluvanna, Texas; (b) the Krasnoyarsk Dam in Russia; (c) a solar farm; (d) a group of nickel metal hydride batteries. The voltage output of each device depends on its construction and load. The voltage output equals emf only if there is no load. (credit a: modification of work by Stig Nygaard; credit b: modification of work by "vadimpl"/Wikimedia Commons; credit c: modification of work by "The tdog"/Wikimedia Commons; credit d: modification of work by "Itrados"/Wikimedia Commons)

If the electromotive force is not a force at all, then what is the emf and what is a source of emf? To answer these questions, consider a simple circuit of a 12-V lamp attached to a 12-V battery, as shown in [link]. The battery can be modeled as a two-terminal device that keeps one terminal at a higher electric potential than the second terminal. The higher electric potential is sometimes called the positive terminal and is labeled with a plus sign. The lower-potential terminal is sometimes called the negative terminal and labeled with a minus sign. This is the source of the emf.


Emf
Source
+
−
$\vec{E}$
I
I
Force due to electric field on the electron
$\vec{F}_E = q\vec{E}$
Direction of motion of electron

A source of emf maintains one terminal at a higher electric potential than the other terminal, acting as a source of current in a circuit.

When the emf source is not connected to the lamp, there is no net flow of charge within the emf source. Once the battery is connected to the lamp, charges flow from one terminal of the battery, through the lamp (causing the lamp to light), and back to the other terminal of the battery. If we consider positive (conventional) current flow, positive charges leave the positive terminal, travel through the lamp, and enter the negative terminal.

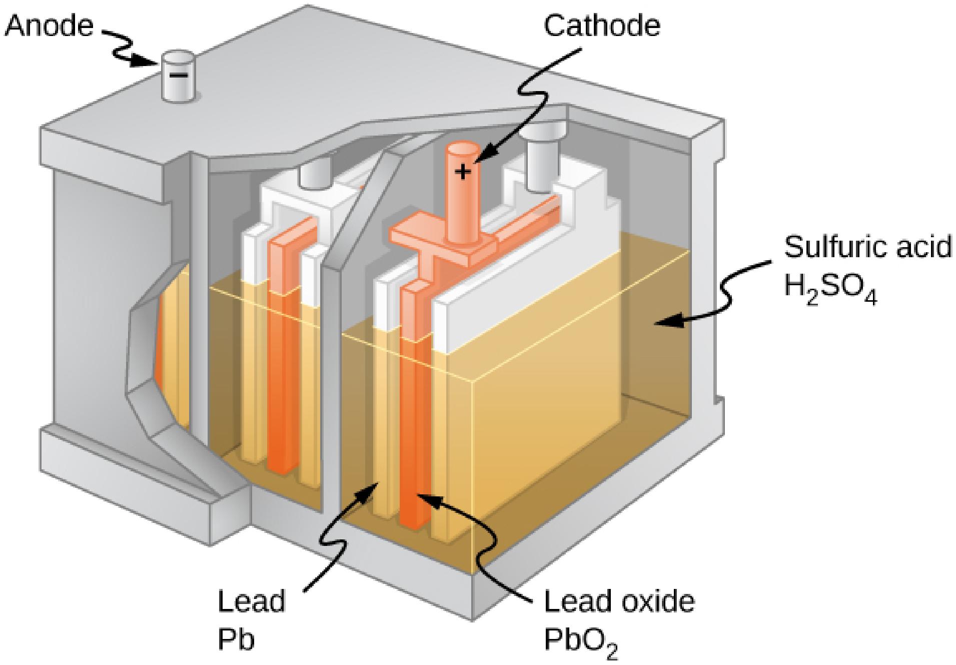
Positive current flow is useful for most of the circuit analysis in this chapter, but in metallic wires and resistors, electrons contribute the most to current, flowing in the opposite direction of positive current flow. Therefore, it is more realistic to consider the movement of electrons for the analysis of the circuit in [link]. The electrons leave the negative terminal, travel through the lamp, and return to the positive terminal. In order for the emf source to maintain the potential difference between the two terminals, negative charges (electrons) must be moved from the positive terminal to the negative terminal. The emf source acts as a charge pump, moving negative charges from the positive terminal to the negative terminal to maintain the potential difference. This increases the potential energy of the charges and, therefore, the electric potential of the charges.

The force on the negative charge from the electric field is in the opposite direction of the electric field, as shown in [link]. In order for the negative charges to be moved to the negative terminal, work must be done on the negative charges. This requires energy, which comes from chemical reactions in the battery. The potential is kept high on the positive terminal and low on the negative terminal to maintain the potential difference between the two terminals. The emf is equal to the work done on the charge per unit charge $\left(\varepsilon = \frac{dW}{dq}\right)$ when there is no current flowing. Since the unit for work is the joule and the unit for charge is the coulomb, the unit for emf is the volt $(1\text{ V} = 1\text{ J/C})$.

The **terminal voltage** V_{terminal} of a battery is voltage measured across the terminals of the battery when there is no load connected to the terminal. An ideal battery is an emf source that maintains a constant terminal voltage, independent of the current between the two terminals. An ideal battery has no internal resistance, and the terminal voltage is equal to the emf of the battery. In the next section, we will show that a real battery does have internal resistance and the terminal voltage is always less than the emf of the battery.

The Origin of Battery Potential

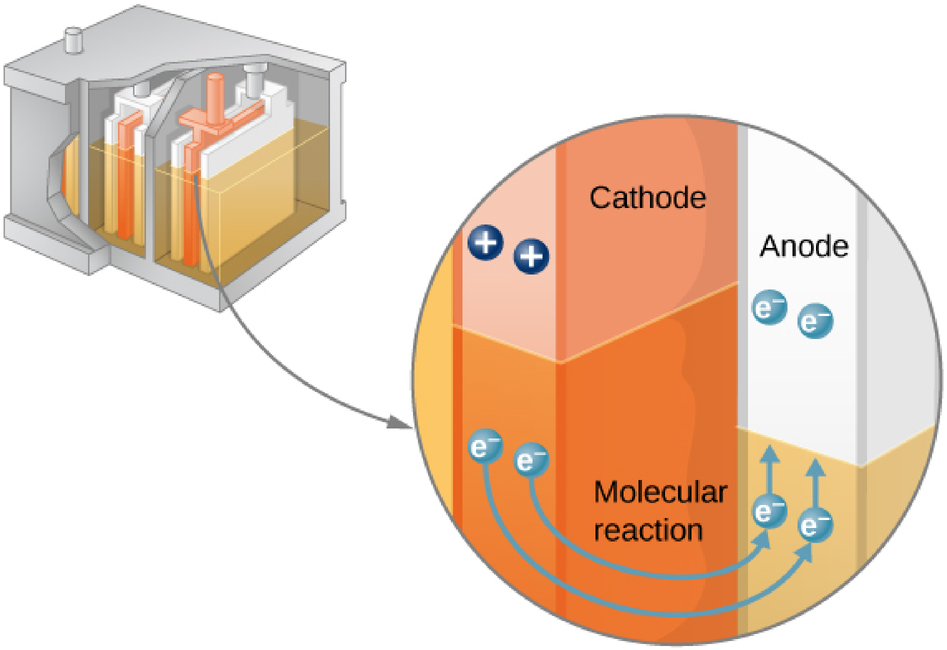
The combination of chemicals and the makeup of the terminals in a battery determine its emf. The lead acid battery used in cars and other vehicles is one of the most common combinations of chemicals. [link] shows a single cell (one of six) of this battery. The cathode (positive) terminal of the cell is connected to a lead oxide plate, whereas the anode (negative) terminal is connected to a lead plate. Both plates are immersed in sulfuric acid, the electrolyte for the system.



Chemical reactions in a lead-acid cell separate charge, sending negative charge to the anode, which is connected to the lead plates. The lead oxide plates are connected to the positive or cathode terminal of the cell. Sulfuric acid conducts the charge, as well as participates in the chemical reaction.

Knowing a little about how the chemicals in a lead-acid battery interact helps in understanding the potential created by the battery. [link] shows the result of a single chemical reaction. Two electrons are placed on the anode, making it negative, provided that the cathode supplies two electrons. This leaves the cathode positively charged, because it has lost two electrons. In short, a separation of charge has been driven by a chemical reaction.

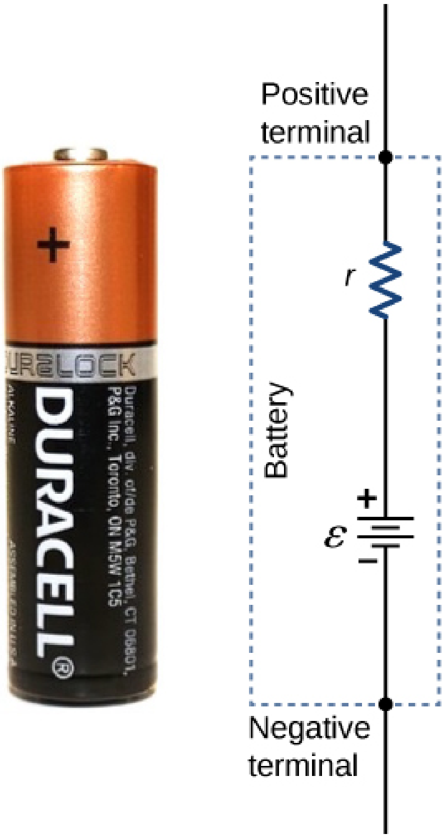
Note that the reaction does not take place unless there is a complete circuit to allow two electrons to be supplied to the cathode. Under many circumstances, these electrons come from the anode, flow through a resistance, and return to the cathode. Note also that since the chemical reactions involve substances with resistance, it is not possible to create the emf without an internal resistance.



In a lead-acid battery, two electrons are forced onto the anode of a cell, and two electrons are removed from the cathode of the cell. The chemical reaction in a lead-acid battery places two electrons on the anode and removes two from the cathode. It requires a closed circuit to proceed, since the two electrons must be supplied to the cathode.

Internal Resistance and Terminal Voltage

The amount of resistance to the flow of current within the voltage source is called the **internal resistance**. The internal resistance r of a battery can behave in complex ways. It generally increases as a battery is depleted, due to the oxidation of the plates or the reduction of the acidity of the electrolyte. However, internal resistance may also depend on the magnitude and direction of the current through a voltage source, its temperature, and even its history. The internal resistance of rechargeable nickel-cadmium cells, for example, depends on how many times and how deeply they have been depleted. A simple model for a battery consists of an idealized emf source ε and an internal resistance r ([link]).



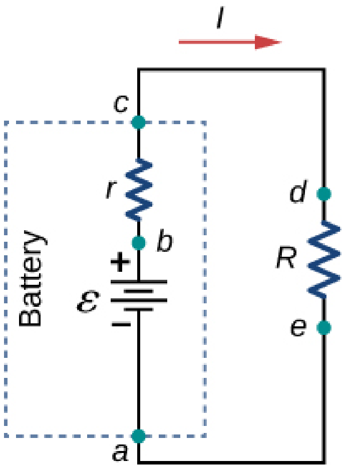
A battery can be modeled as an idealized emf (ε) with an internal resistance (r). The terminal voltage of the battery is $V_{\text{terminal}} = \varepsilon - Ir$.

Suppose an external resistor, known as the load resistance R, is connected to a voltage source such as a battery, as in [link]. The figure shows a model of a battery with an emf ε, an internal resistance r, and a load resistor R connected across its terminals. Using conventional current flow, positive charges leave the positive terminal of the battery, travel through the resistor, and return to the negative terminal of the battery. The terminal voltage of the battery depends on the emf, the internal resistance, and the current, and is equal to

Note:
Equation:

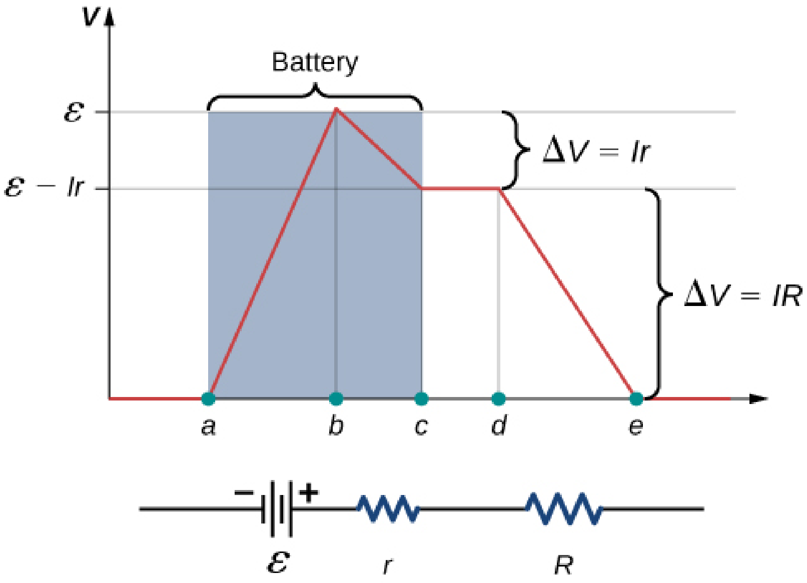
$$V_{\text{terminal}} = \varepsilon - Ir.$$

For a given emf and internal resistance, the terminal voltage decreases as the current increases due to the potential drop Ir of the internal resistance.



Schematic of a voltage source and its load resistor R. Since the internal resistance r is in series with the load, it can significantly affect the terminal voltage and the current delivered to the load.

A graph of the potential difference across each element the circuit is shown in [link]. A current I runs through the circuit, and the potential drop across the internal resistor is equal to Ir. The terminal voltage is equal to $\varepsilon - Ir$, which is equal to the **potential drop** across the load resistor $IR = \varepsilon - Ir$. As with potential energy, it is the change in voltage that is important. When the term "voltage" is used, we assume that it is actually the change in the potential, or ΔV. However, Δ is often omitted for convenience.



A graph of the voltage through the circuit of a battery and a load resistance. The electric potential increases the emf of the battery due to the chemical reactions doing work on the charges. There is a decrease in the electric potential in the battery due to the internal resistance. The potential decreases due to the internal resistance $(-Ir)$, making the terminal voltage of the battery equal to $(\varepsilon - Ir)$. The voltage then decreases by (IR). The current is equal to $I = \frac{\varepsilon}{r+R}$.

The current through the load resistor is $I = \frac{\varepsilon}{r+R}$. We see from this expression that the smaller the internal resistance r, the greater the current the voltage source supplies to its load R. As batteries are depleted, r increases. If r becomes a significant fraction of the load resistance, then the current is significantly reduced, as the following example illustrates.

Example:

Analyzing a Circuit with a Battery and a Load

A given battery has a 12.00-V emf and an internal resistance of $0.100\ \Omega$. (a) Calculate its terminal voltage when connected to a 10.00-Ω load. (b) What is the terminal voltage when connected to a 0.500-Ω load? (c) What power does the 0.500-Ω load dissipate? (d) If the internal resistance grows to $0.500\ \Omega$, find the current, terminal voltage, and power dissipated by a 0.500-Ω load.

Strategy

The analysis above gave an expression for current when internal resistance is taken into account. Once the current is found, the terminal voltage can be calculated by using the equation $V_{\text{terminal}} = \varepsilon - Ir$. Once current is found, we can also find the power dissipated by the resistor.

Solution

a. Entering the given values for the emf, load resistance, and internal resistance into the expression above yields
   Equation:

$$I = \frac{\varepsilon}{R + r} = \frac{12.00\ \text{V}}{10.10\ \Omega} = 1.188\ \text{A}.$$

   Enter the known values into the equation $V_{\text{terminal}} = \varepsilon - Ir$ to get the terminal voltage:
   Equation:

$$V_{\text{terminal}} = \varepsilon - Ir = 12.00\ \text{V}\ - (1.188\ \text{A})(0.100\ \Omega) = 11.90\ \text{V}.$$

   The terminal voltage here is only slightly lower than the emf, implying that the current drawn by this light load is not significant.

b. Similarly, with $R_{\text{load}} = 0.500\ \Omega$, the current is
   Equation:

$$I = \frac{\varepsilon}{R + r} = \frac{12.00\ \text{V}}{0.600\ \Omega} = 20.00\ \text{A}.$$

   The terminal voltage is now
   Equation:

$$V_{\text{terminal}} = \varepsilon - Ir = 12.00\ \text{V} - (20.00\ \text{A})\,(0.100\ \Omega) = 10.00\ \text{V}.$$

   The terminal voltage exhibits a more significant reduction compared with emf, implying $0.500\ \Omega$ is a heavy load for this battery. A “heavy load” signifies a larger draw of current from the source but not a larger resistance.

c. The power dissipated by the $0.500\text{-}\Omega$ load can be found using the formula $P = I^2R$. Entering the known values gives
Equation:

$$P = I^2R = (20.0 \text{ A})^2(0.500 \ \Omega) = 2.00 \ \times \ 10^2 \text{ W}.$$

Note that this power can also be obtained using the expression $\frac{V^2}{R}$ or IV, where V is the terminal voltage (10.0 V in this case).

d. Here, the internal resistance has increased, perhaps due to the depletion of the battery, to the point where it is as great as the load resistance. As before, we first find the current by entering the known values into the expression, yielding
Equation:

$$I = \frac{\varepsilon}{R + r} = \frac{12.00 \text{ V}}{1.00 \ \Omega} = 12.00 \text{ A}.$$

Now the terminal voltage is
Equation:

$$V_{\text{terminal}} = \varepsilon - Ir = 12.00 \text{ V} - (12.00 \text{ A})(0.500 \ \Omega) = 6.00 \text{ V},$$

and the power dissipated by the load is
Equation:

$$P = I^2R = (12.00 \text{ A})^2(0.500 \ \Omega) = 72.00 \text{ W}.$$

We see that the increased internal resistance has significantly decreased the terminal voltage, current, and power delivered to a load.

Significance
The internal resistance of a battery can increase for many reasons. For example, the internal resistance of a rechargeable battery increases as the number of times the battery is recharged increases. The increased internal resistance may have two effects on the battery. First, the terminal voltage will decrease. Second, the battery may overheat due to the increased power dissipated by the internal resistance.

Note:
Exercise:

Problem:

Check Your Understanding If you place a wire directly across the two terminal of a battery, effectively shorting out the terminals, the battery will begin to get hot. Why do you suppose this happens?

Solution:

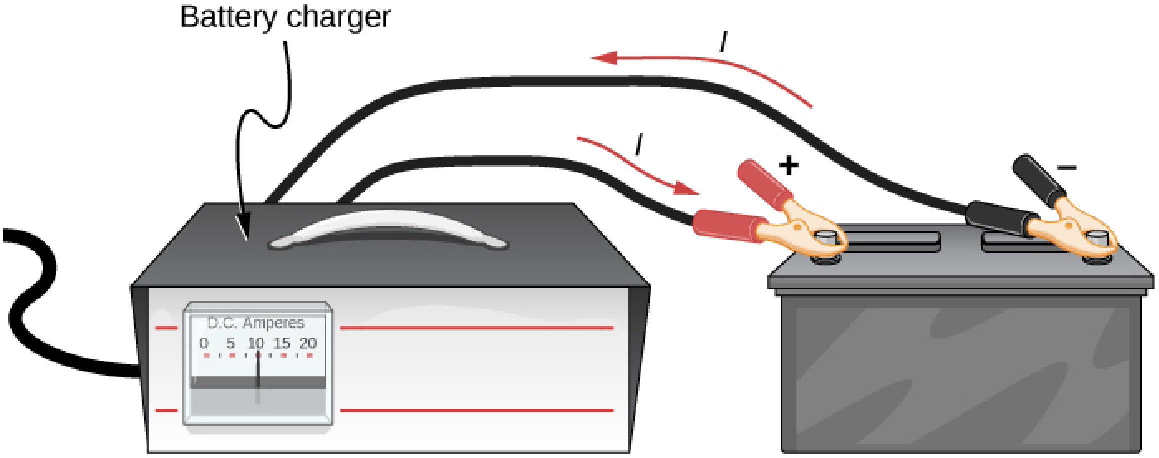
If a wire is connected across the terminals, the load resistance is close to zero, or at least considerably less than the internal resistance of the battery. Since the internal resistance is small, the current through the circuit will be large, $I=\frac{\varepsilon}{R+r}=\frac{\varepsilon}{0+r}=\frac{\varepsilon}{r}$. The large current causes a high power to be dissipated by the internal resistance $\left(P=I^2r\right)$. The power is dissipated as heat.

Battery Testers

Battery testers, such as those in [link], use small load resistors to intentionally draw current to determine whether the terminal potential drops below an acceptable level. Although it is difficult to measure the internal resistance of a battery, battery testers can provide a measurement of the internal resistance of the battery. If internal resistance is high, the battery is weak, as evidenced by its low terminal voltage.

Battery testers measure terminal voltage under a load to determine the condition of a battery. (a) A US Navy electronics technician uses a battery tester to test large batteries aboard the aircraft carrier USS *Nimitz*. The battery tester she uses has a small resistance that can dissipate large amounts of power. (b) The small device shown is used on small batteries and has a digital display to indicate the acceptability of the terminal voltage. (credit a: modification of work by Jason A. Johnston; credit b: modification of work by Keith Williamson)

Some batteries can be recharged by passing a current through them in the direction opposite to the current they supply to an appliance. This is done routinely in cars and in batteries for small electrical appliances and electronic devices ([link]). The voltage output of the battery charger must be greater than the emf of the battery to reverse the current through it. This causes the terminal voltage of the battery to be greater than the emf, since $V = \varepsilon - Ir$ and I is now negative.



A car battery charger reverses the normal direction of current through a battery, reversing its chemical reaction and replenishing its chemical potential.

It is important to understand the consequences of the internal resistance of emf sources, such as batteries and solar cells, but often, the analysis of circuits is done with the terminal voltage of the battery, as we have done in the previous sections. The terminal voltage is referred to as simply as V, dropping the subscript "terminal." This is because the internal resistance of the battery is difficult to measure directly and can change over time.

Summary

- All voltage sources have two fundamental parts: a source of electrical energy that has a characteristic electromotive force (emf), and an internal resistance r. The emf is the work done per charge to keep the potential difference of a source constant. The emf is equal to the potential difference across the terminals when no current is flowing. The

internal resistance *r* of a voltage source affects the output voltage when a current flows.
- The voltage output of a device is called its terminal voltage V_{terminal} and is given by $V_{\text{terminal}} = \varepsilon - Ir$, where *I* is the electric current and is positive when flowing away from the positive terminal of the voltage source and *r* is the internal resistance.

Conceptual Questions

Exercise:

Problem:

What effect will the internal resistance of a rechargeable battery have on the energy being used to recharge the battery?

Solution:

Some of the energy being used to recharge the battery will be dissipated as heat by the internal resistance.

Exercise:

Problem:

A battery with an internal resistance of *r* and an emf of 10.00 V is connected to a load resistor $R = r$. As the battery ages, the internal resistance triples. How much is the current through the load resistor reduced?

Exercise:

Problem:

Show that the power dissipated by the load resistor is maximum when the resistance of the load resistor is equal to the internal resistance of the battery.

Solution:

$$P = I^2R = \left(\tfrac{\varepsilon}{r+R}\right)^2 R = \varepsilon^2 R(r+R)^{-2}, \;\; \tfrac{dP}{dR} = \varepsilon^2\left[(r+R)^{-2} - 2R(r+R)^{-3}\right] = 0,$$
$$\left[\tfrac{(r+R)-2R}{(r+R)^3}\right] = 0, \;\; r = R$$

Problems

Exercise:

Problem:

A car battery with a 12-V emf and an internal resistance of $0.050\ \Omega$ is being charged with a current of 60 A. Note that in this process, the battery is being charged. (a) What is the potential difference across its terminals? (b) At what rate is thermal energy being dissipated in the battery? (c) At what rate is electric energy being converted into chemical energy?

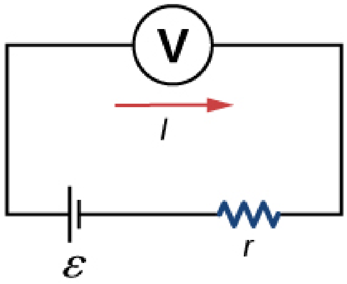
Exercise:

Problem:

The label on a battery-powered radio recommends the use of a rechargeable nickel-cadmium cell (nicads), although it has a 1.25-V emf, whereas an alkaline cell has a 1.58-V emf. The radio has a $3.20\ \Omega$ resistance. (a) Draw a circuit diagram of the radio and its battery. Now, calculate the power delivered to the radio (b) when using a nicad cells, each having an internal resistance of $0.0400\ \Omega$, and (c) when using an alkaline cell, having an internal resistance of $0.200\ \Omega$. (d) Does this difference seem significant, considering that the radio's effective resistance is lowered when its volume is turned up?

Solution:

a.



b. 0.476W; c. 0.691 W; d. As R_L is lowered, the power difference decreases; therefore, at higher volumes, there is no significant difference.

Exercise:

Problem:

An automobile starter motor has an equivalent resistance of $0.0500\ \Omega$ and is supplied by a 12.0-V battery with a $0.0100\text{-}\Omega$ internal resistance. (a) What is the current to the motor? (b) What voltage is applied to it? (c) What power is supplied to the motor? (d) Repeat these calculations for when the battery connections are corroded and add $0.0900\ \Omega$ to the circuit. (Significant problems are caused by even small amounts of unwanted resistance in low-voltage, high-current applications.)

Exercise:

Problem:

(a) What is the internal resistance of a voltage source if its terminal potential drops by 2.00 V when the current supplied increases by 5.00 A? (b) Can the emf of the voltage source be found with the information supplied?

Solution:

a. $0.400\ \Omega$; b. No, there is only one independent equation, so only *r* can be found.

Exercise:

Problem:

A person with body resistance between his hands of $10.0\ \text{k}\Omega$ accidentally grasps the terminals of a 20.0-kV power supply. (Do NOT do this!) (a) Draw a circuit diagram to represent the situation. (b) If the internal resistance of the power supply is $2000\ \Omega$, what is the current through his body? (c) What is the power dissipated in his body? (d) If the power supply is to be made safe by increasing its internal resistance, what should the internal resistance be for the maximum current in this situation to be 1.00 mA or less? (e) Will this modification compromise the effectiveness of the power supply for driving low-resistance devices? Explain your reasoning.

Exercise:

Problem:

A 12.0-V emf automobile battery has a terminal voltage of 16.0 V when being charged by a current of 10.0 A. (a) What is the battery's internal resistance? (b) What power is dissipated inside the battery? (c) At what rate (in $^\circ\text{C/min}$) will its temperature increase if its mass is 20.0 kg and it has a specific heat of $0.300\ \text{kcal/kg}\ \cdot\ ^\circ\text{C}$, assuming no heat escapes?

Solution:

a. $0.400\ \Omega$; b. 40.0 W; c. $0.0956\ ^\circ\text{C/min}$

Glossary

electromotive force (emf)
: energy produced per unit charge, drawn from a source that produces an electrical current

internal resistance
: amount of resistance to the flow of current within the voltage source

potential difference

difference in electric potential between two points in an electric circuit, measured in volts

potential drop

loss of electric potential energy as a current travels across a resistor, wire, or other component

terminal voltage

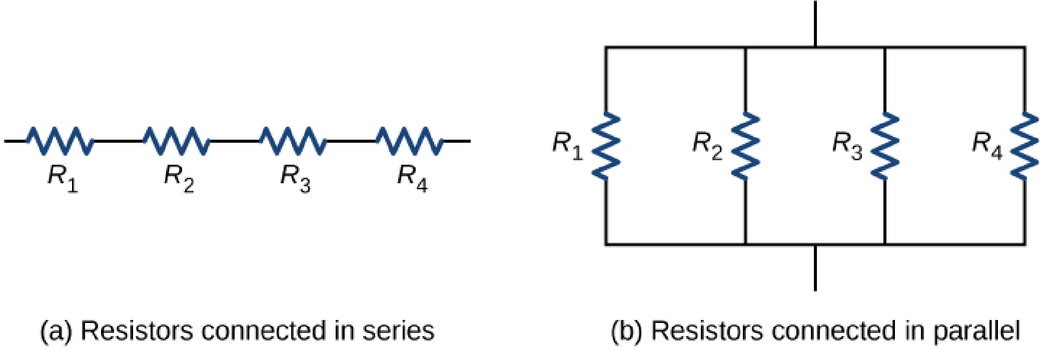
potential difference measured across the terminals of a source when there is no load attached

Resistors in Series and Parallel
By the end of the section, you will be able to:

- Define the term equivalent resistance
- Calculate the equivalent resistance of resistors connected in series
- Calculate the equivalent resistance of resistors connected in parallel

In Current and Resistance, we described the term ‘resistance’ and explained the basic design of a resistor. Basically, a resistor limits the flow of charge in a circuit and is an ohmic device where $V = IR$. Most circuits have more than one resistor. If several resistors are connected together and connected to a battery, the current supplied by the battery depends on the **equivalent resistance** of the circuit.

The equivalent resistance of a combination of resistors depends on both their individual values and how they are connected. The simplest combinations of resistors are series and parallel connections ([link]). In a series circuit, the output current of the first resistor flows into the input of the second resistor; therefore, the current is the same in each resistor. In a parallel circuit, all of the resistor leads on one side of the resistors are connected together and all the leads on the other side are connected together. In the case of a parallel configuration, each resistor has the same potential drop across it, and the currents through each resistor may be different, depending on the resistor. The sum of the individual currents equals the current that flows into the parallel connections.

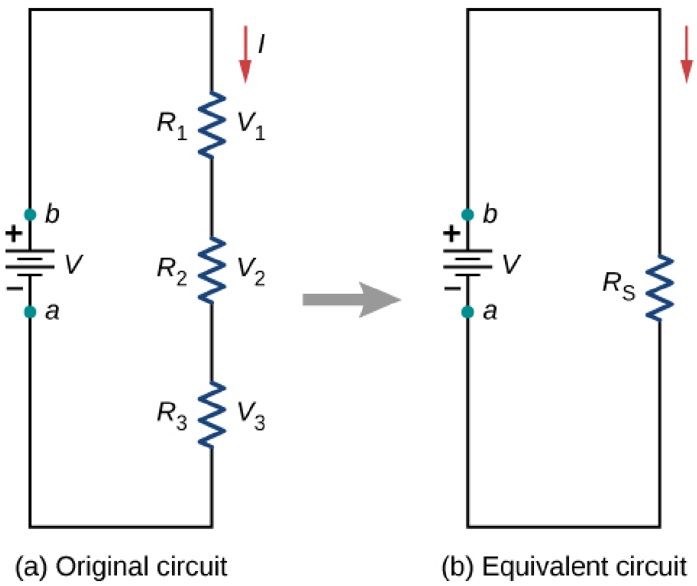


(a) Resistors connected in series (b) Resistors connected in parallel

(a) For a series connection of resistors, the current is the same in each resistor. (b) For a parallel connection of resistors, the voltage is the same across each resistor.

Resistors in Series

Resistors are said to be in series whenever the current flows through the resistors sequentially. Consider [link], which shows three resistors in series with an applied voltage equal to V_{ab}. Since there is only one path for the charges to flow through, the current is the same through each resistor. The equivalent resistance of a set of resistors in a series connection is equal to the algebraic sum of the individual resistances.



(a) Three resistors connected in series to a voltage source. (b) The original circuit is reduced to an equivalent resistance and a voltage source.

In [link], the current coming from the voltage source flows through each resistor, so the current through each resistor is the same. The current through the circuit depends on the voltage supplied by the voltage source and the resistance of the resistors. For each resistor, a potential drop occurs that is equal to the loss of electric potential energy as a current travels through each resistor. According to Ohm's law, the potential drop V across a resistor when a current flows through it is calculated using the equation $V=IR$, where I is the current in amps (A) and R is the resistance in ohms (Ω). Since energy is conserved, and the voltage is equal to the potential energy per charge, the sum of the voltage applied to the circuit by the source and the potential drops across the individual resistors around a loop should be equal to zero:

Equation:

$$\sum_{i=1}^{N} V_i = 0.$$

This equation is often referred to as Kirchhoff's loop law, which we will look at in more detail later in this chapter. For [link], the sum of the potential drop of each resistor and the voltage supplied by the voltage source should equal zero:

Equation:

$$\begin{aligned} V - V_1 - V_2 - V_3 &= 0, \\ V &= V_1 + V_2 + V_3, \\ &= IR_1 + IR_2 + IR_3, \\ I &= \frac{V}{R_1+R_2+R_3} = \frac{V}{R_S}. \end{aligned}$$

Since the current through each component is the same, the equality can be simplified to an equivalent resistance, which is just the sum of the resistances of the individual resistors.

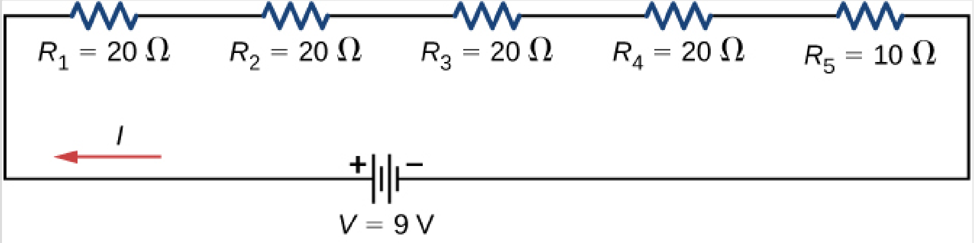
Any number of resistors can be connected in series. If N resistors are connected in series, the equivalent resistance is

Note:
Equation:

$$R_S = R_1 + R_2 + R_3 + \cdots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i.$$

One result of components connected in a series circuit is that if something happens to one component, it affects all the other components. For example, if several lamps are connected in series and one bulb burns out, all the other lamps go dark.

Example:
Equivalent Resistance, Current, and Power in a Series Circuit
A battery with a terminal voltage of 9 V is connected to a circuit consisting of four 20-Ω and one 10-Ω resistors all in series ([link]). Assume the battery has negligible internal resistance. (a) Calculate the equivalent resistance of the circuit. (b) Calculate the current through each resistor. (c) Calculate the potential drop across each resistor. (d) Determine the total power dissipated by the resistors and the power supplied by the battery.



A simple series circuit with five resistors.

Strategy
In a series circuit, the equivalent resistance is the algebraic sum of the resistances. The current through the circuit can be found from Ohm’s law and is equal to the voltage divided by the equivalent resistance. The potential drop across each resistor can be found using Ohm’s law. The power dissipated by each resistor can be found using $P = I^2R$, and the total power dissipated by the resistors is equal to the sum of the power dissipated by each resistor. The power supplied by the battery can be found using $P = I\varepsilon$.
Solution

a. The equivalent resistance is the algebraic sum of the resistances:
   Equation:

$$R_S = R_1 + R_2 + R_3 + R_4 + R_5 = 20\,\Omega + 20\,\Omega + 20\,\Omega + 20\,\Omega + 10\,\Omega = 90\,\Omega.$$

b. The current through the circuit is the same for each resistor in a series circuit and is equal to the applied voltage divided by the equivalent resistance:
Equation:

$$I = \frac{V}{R_S} = \frac{9\text{ V}}{90\,\Omega} = 0.1\text{ A}.$$

c. The potential drop across each resistor can be found using Ohm's law:
Equation:

$$\begin{aligned} &V_1 = V_2 = V_3 = V_4 = (0.1\text{ A})20\,\Omega = 2\text{ V}, \\ &V_5 = (0.1\text{ A})10\,\Omega = 1\text{ V}, \\ &V_1 + V_2 + V_3 + V_4 + V_5 = 9\text{ V}. \end{aligned}$$

Note that the sum of the potential drops across each resistor is equal to the voltage supplied by the battery.

d. The power dissipated by a resistor is equal to $P = I^2R$, and the power supplied by the battery is equal to $P = I\varepsilon$:
Equation:

$$\begin{aligned} &P_1 = P_2 = P_3 = P_4 = (0.1\text{ A})^2\,(20\,\Omega) = 0.2\text{ W}, \\ &P_5 = (0.1\text{ A})^2\,(10\,\Omega) = 0.1\text{ W}, \\ &P_{\text{dissipated}} = 0.2\text{ W} + 0.2\text{ W} + 0.2\text{ W} + 0.2\text{ W} + 0.1\text{ W} = 0.9\text{ W}, \\ &P_{\text{source}} = I\varepsilon = (0.1\text{ A})\,(9\text{ V}) = 0.9\text{ W}. \end{aligned}$$

Significance

There are several reasons why we would use multiple resistors instead of just one resistor with a resistance equal to the equivalent resistance of the circuit. Perhaps a resistor of the required size is not available, or we need to dissipate the heat generated, or we want to minimize the cost of resistors. Each resistor may cost a few cents to a few dollars, but when multiplied by thousands of units, the cost saving may be appreciable.

Note:
Exercise:

Problem:

Check Your Understanding Some strings of miniature holiday lights are made to short out when a bulb burns out. The device that causes the short is called a shunt, which allows current to flow around the open circuit. A "short" is like putting a piece of wire across the component. The bulbs are usually grouped in series of nine bulbs. If too many bulbs burn out, the shunts eventually open. What causes this?

Solution:

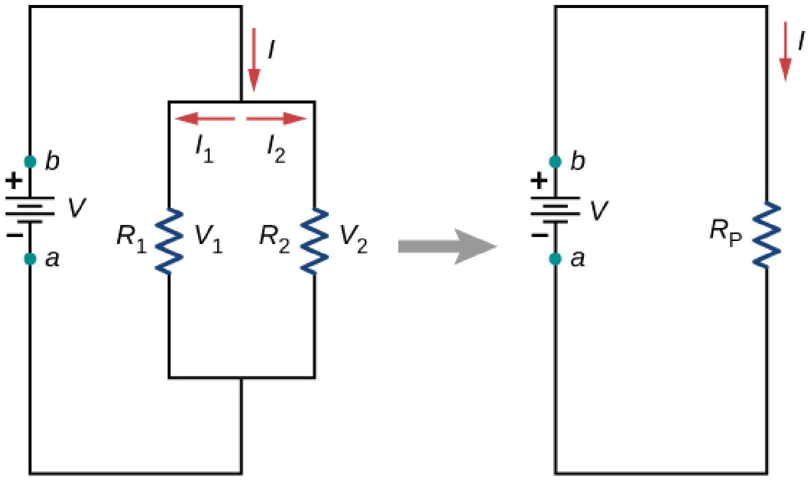
> The equivalent resistance of nine bulbs connected in series is $9R$. The current is $I = V/9\,R$. If one bulb burns out, the equivalent resistance is $8R$, and the voltage does not change, but the current increases $(I = V/8\,R)$. As more bulbs burn out, the current becomes even higher. Eventually, the current becomes too high, burning out the shunt.

Let’s briefly summarize the major features of resistors in series:

1. Series resistances add together to get the equivalent resistance:
   Equation:
   $$R_{\text{S}} = R_1 + R_2 + R_3 + \cdots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i.$$
2. The same current flows through each resistor in series.
3. Individual resistors in series do not get the total source voltage, but divide it. The total potential drop across a series configuration of resistors is equal to the sum of the potential drops across each resistor.

Resistors in Parallel

[link] shows resistors in parallel, wired to a voltage source. Resistors are in parallel when one end of all the resistors are connected by a continuous wire of negligible resistance and the other end of all the resistors are also connected to one another through a continuous wire of negligible resistance. The potential drop across each resistor is the same. Current through each resistor can be found using Ohm’s law $I = V/R$, where the voltage is constant across each resistor. For example, an automobile’s headlights, radio, and other systems are wired in parallel, so that each subsystem utilizes the full voltage of the source and can operate completely independently. The same is true of the wiring in your house or any building.



(a) Original circuit

(b) Equivalent circuit

(a) Two resistors connected in parallel to a voltage source. (b) The original circuit is reduced to an equivalent resistance and a

The original circuit is reduced to an equivalent resistance and a voltage source.

The current flowing from the voltage source in [link] depends on the voltage supplied by the voltage source and the equivalent resistance of the circuit. In this case, the current flows from the voltage source and enters a junction, or node, where the circuit splits flowing through resistors R_1 and R_2. As the charges flow from the battery, some go through resistor R_1 and some flow through resistor R_2. The sum of the currents flowing into a junction must be equal to the sum of the currents flowing out of the junction:

Equation:

$$\sum I_{\text{in}} = \sum I_{\text{out}}.$$

This equation is referred to as Kirchhoff's junction rule and will be discussed in detail in the next section. In [link], the junction rule gives $I = I_1 + I_2$. There are two loops in this circuit, which leads to the equations $V = I_1 R_1$ and $I_1 R_1 = I_2 R_2$. Note the voltage across the resistors in parallel are the same $(V = V_1 = V_2)$ and the current is additive:

Equation:

$$\begin{aligned} I &= I_1 + I_2 \\ &= \tfrac{V_1}{R_1} + \tfrac{V_2}{R_2} \\ &= \tfrac{V}{R_1} + \tfrac{V}{R_2} \\ &= V\left(\tfrac{1}{R_1} + \tfrac{1}{R_2}\right) = \tfrac{V}{R_{\text{P}}} \\ R_{\text{P}} &= \left(\tfrac{1}{R_1} + \tfrac{1}{R_2}\right)^{-1}. \end{aligned}$$

Generalizing to any number of *N* resistors, the equivalent resistance R_{P} of a parallel connection is related to the individual resistances by

Note:

Equation:

$$R_{\text{P}} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots + \frac{1}{R_{N-1}} + \frac{1}{R_N}\right)^{-1} = \left(\sum_{i=1}^{N} \frac{1}{R_i}\right)^{-1}.$$

This relationship results in an equivalent resistance R_{P} that is less than the smallest of the individual resistances. When resistors are connected in parallel, more current flows from the source than would flow for any of them individually, so the total resistance is lower.

Example:

Analysis of a Parallel Circuit

Three resistors $R_1 = 1.00\ \Omega$, $R_2 = 2.00\ \Omega$, and $R_3 = 2.00\ \Omega$, are connected in parallel. The parallel connection is attached to a $V = 3.00\ \text{V}$ voltage source. (a) What is the equivalent resistance? (b) Find the current supplied by the source to the parallel circuit. (c) Calculate the currents in each resistor and show that these add together to equal the current output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source and show that it equals the total power dissipated by the resistors.

Strategy

(a) The total resistance for a parallel combination of resistors is found using $R_\text{P} = \left(\sum_i \frac{1}{R_i}\right)^{-1}$.

(Note that in these calculations, each intermediate answer is shown with an extra digit.)

(b) The current supplied by the source can be found from Ohm's law, substituting R_P for the total resistance $I = \frac{V}{R_\text{P}}$.

(c) The individual currents are easily calculated from Ohm's law $\left(I_i = \frac{V_i}{R_i}\right)$, since each resistor gets the full voltage. The total current is the sum of the individual currents: $I = \sum_i I_i$.

(d) The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P_i = V^2/R_i$, since each resistor gets full voltage.

(e) The total power can also be calculated in several ways, use $P = IV$.

Solution

a. The total resistance for a parallel combination of resistors is found using [link]. Entering known values gives
   Equation:

$$R_\text{P} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)^{-1} = \left(\frac{1}{1.00\ \Omega} + \frac{1}{2.00\ \Omega} + \frac{1}{2.00\ \Omega}\right)^{-1} = 0.50\ \Omega.$$

   The total resistance with the correct number of significant digits is $R_\text{P} = 0.50\ \Omega$. As predicted, R_P is less than the smallest individual resistance.

b. The total current can be found from Ohm's law, substituting R_P for the total resistance. This gives
   Equation:

$$I = \frac{V}{R_\text{P}} = \frac{3.00\ \text{V}}{0.50\ \Omega} = 6.00\ \text{A}.$$

   Current *I* for each device is much larger than for the same devices connected in series (see the previous example). A circuit with parallel connections has a smaller total resistance than the resistors connected in series.

c. The individual currents are easily calculated from Ohm's law, since each resistor gets the full voltage. Thus,
   Equation:

$$I_1 = \frac{V}{R_1} = \frac{3.00\ \text{V}}{1.00\ \Omega} = 3.00\ \text{A}.$$

Similarly,
Equation:

$$I_2 = \frac{V}{R_2} = \frac{3.00\text{ V}}{2.00\,\Omega} = 1.50\text{ A}$$

and
Equation:

$$I_3 = \frac{V}{R_3} = \frac{6.00\text{ V}}{2.00\,\Omega} = 1.50\text{ A.}$$

The total current is the sum of the individual currents:
Equation:

$$I_1 + I_2 + I_3 = 6.00\text{ A.}$$

d. The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P = V^2/R$, since each resistor gets full voltage. Thus,
Equation:

$$P_1 = \frac{V^2}{R_1} = \frac{(3.00\text{ V})^2}{1.00\,\Omega} = 9.00\text{ W.}$$

Similarly,
Equation:

$$P_2 = \frac{V^2}{R_2} = \frac{(3.00\text{ V})^2}{2.00\,\Omega} = 4.50\text{ W}$$

and
Equation:

$$P_3 = \frac{V^2}{R_3} = \frac{(3.00\text{ V})^2}{2.00\,\Omega} = 4.50\text{ W.}$$

e. The total power can also be calculated in several ways. Choosing $P = IV$ and entering the total current yields
Equation:

$$P = IV = (6.00\text{ A})(3.00\text{ V}) = 18.00\text{ W.}$$

Significance
Total power dissipated by the resistors is also 18.00 W:
Equation:

$$P_1 + P_2 + P_3 = 9.00\ \text{W} + 4.50\ \text{W} + 4.50\ \text{W} = 18.00\ \text{W}.$$

Notice that the total power dissipated by the resistors equals the power supplied by the source.

Note:
Exercise:

Problem:

Check Your Understanding Consider the same potential difference $(V = 3.00\ \text{V})$ applied to the same three resistors connected in series. Would the equivalent resistance of the series circuit be higher, lower, or equal to the three resistor in parallel? Would the current through the series circuit be higher, lower, or equal to the current provided by the same voltage applied to the parallel circuit? How would the power dissipated by the resistor in series compare to the power dissipated by the resistors in parallel?

Solution:

The equivalent of the series circuit would be $R_{\text{eq}} = 1.00\ \Omega + 2.00\ \Omega + 2.00\ \Omega = 5.00\ \Omega$, which is higher than the equivalent resistance of the parallel circuit $R_{\text{eq}} = 0.50\ \Omega$. The equivalent resistor of any number of resistors is always higher than the equivalent resistance of the same resistors connected in parallel. The current through for the series circuit would be $I = \frac{3.00\ \text{V}}{5.00\ \Omega} = 0.60\ \text{A}$, which is lower than the sum of the currents through each resistor in the parallel circuit, $I = 6.00\ \text{A}$. This is not surprising since the equivalent resistance of the series circuit is higher. The current through a series connection of any number of resistors will always be lower than the current into a parallel connection of the same resistors, since the equivalent resistance of the series circuit will be higher than the parallel circuit. The power dissipated by the resistors in series would be $P = 1.80\ \text{W}$, which is lower than the power dissipated in the parallel circuit $P = 18.00\ \text{W}$.

Note:
Exercise:

Problem:

Check Your Understanding How would you use a river and two waterfalls to model a parallel configuration of two resistors? How does this analogy break down?

Solution:

A river, flowing horizontally at a constant rate, splits in two and flows over two waterfalls. The water molecules are analogous to the electrons in the parallel circuits. The number of water molecules that flow in the river and falls must be equal to the number of molecules that flow over each waterfall, just like sum of the current through each resistor must be equal to the current flowing into the parallel circuit. The water molecules in the river have energy due to their motion and height. The potential energy of the water molecules in the river is constant due to their equal heights. This is analogous to the constant change in voltage across a parallel circuit. Voltage is the potential energy across each resistor.
The analogy quickly breaks down when considering the energy. In the waterfall, the potential energy is converted into kinetic energy of the water molecules. In the case of electrons flowing

> through a resistor, the potential drop is converted into heat and light, not into the kinetic energy of the electrons.

Let us summarize the major features of resistors in parallel:

1. Equivalent resistance is found from
   Equation:
   $$R_\text{P} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots + \frac{1}{R_{N-1}} + \frac{1}{R_N}\right)^{-1} = \left(\sum_{i=1}^{N} \frac{1}{R_i}\right)^{-1},$$
   and is smaller than any individual resistance in the combination.
2. The potential drop across each resistor in parallel is the same.
3. Parallel resistors do not each get the total current; they divide it. The current entering a parallel combination of resistors is equal to the sum of the current through each resistor in parallel.

In this chapter, we introduced the equivalent resistance of resistors connect in series and resistors connected in parallel. You may recall that in Capacitance, we introduced the equivalent capacitance of capacitors connected in series and parallel. Circuits often contain both capacitors and resistors. [link] summarizes the equations used for the equivalent resistance and equivalent capacitance for series and parallel connections.

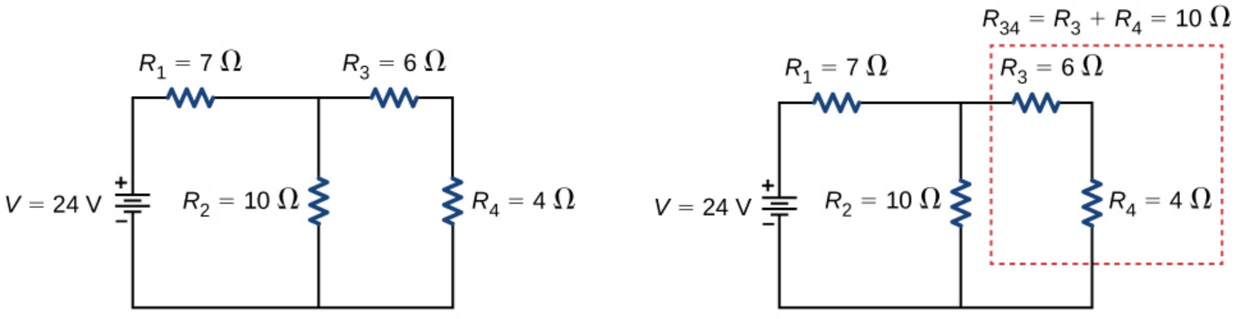
	Series combination	**Parallel combination**
Equivalent capacitance	$\frac{1}{C_\text{S}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots$	$C_\text{P} = C_1 + C_2 + C_3 + \cdots$
Equivalent resistance	$R_\text{S} = R_1 + R_2 + R_3 + \cdots = \sum_{i=1}^{N} R_i$	$\frac{1}{R_\text{P}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots$

Summary for Equivalent Resistance and Capacitance in Series and Parallel Combinations

Combinations of Series and Parallel

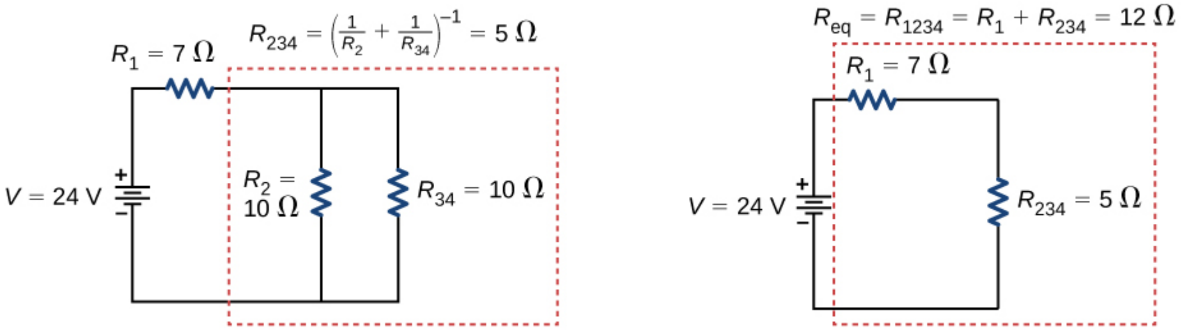
More complex connections of resistors are often just combinations of series and parallel connections. Such combinations are common, especially when wire resistance is considered. In that case, wire resistance is in series with other resistances that are in parallel.

Combinations of series and parallel can be reduced to a single equivalent resistance using the technique illustrated in [link]. Various parts can be identified as either series or parallel connections, reduced to their equivalent resistances, and then further reduced until a single equivalent resistance is left. The process is more time consuming than difficult. Here, we note the equivalent resistance as R_eq.


$R_1 = 7\,\Omega$
$R_3 = 6\,\Omega$
$V = 24$ V
+
−
$R_2 = 10\,\Omega$
$R_4 = 4\,\Omega$
$R_{34} = R_3 + R_4 = 10\,\Omega$
$R_1 = 7\,\Omega$
$R_3 = 6\,\Omega$
$V = 24$ V
+
−
$R_2 = 10\,\Omega$
$R_4 = 4\,\Omega$

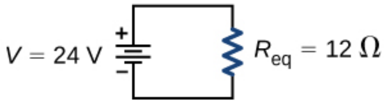
(a) Circuit schematic

(b) Step 1: resistors R_3 and R_4 in series


$R_{234} = \left(\frac{1}{R_2} + \frac{1}{R_{34}}\right)^{-1} = 5\,\Omega$
$R_1 = 7\,\Omega$
$V = 24$ V
+
−
$R_2 = 10\,\Omega$
$R_{34} = 10\,\Omega$
$R_{eq} = R_{1234} = R_1 + R_{234} = 12\,\Omega$
$R_1 = 7\,\Omega$
$V = 24$ V
+
−
$R_{234} = 5\,\Omega$

(c) Step 2: resistors R_2 and R_{34} in parallel

(d) Step 3: resistors R_1 and R_{234} in series


$V = 24$ V
+
−
$R_{eq} = 12\,\Omega$

(e) Simplified schematic reflecting equivalent resistance R_{eq}

(a) The original circuit of four resistors. (b) Step 1: The resistors R_3 and R_4 are in series and the equivalent resistance is $R_{34} = 10\,\Omega$. (c) Step 2: The reduced circuit shows resistors R_2 and R_{34} are in parallel, with an equivalent resistance of $R_{234} = 5\,\Omega$. (d) Step 3: The reduced circuit shows that R_1 and R_{234} are in series with an equivalent resistance of $R_{1234} = 12\,\Omega$, which is the equivalent resistance R_{eq}. (e) The reduced circuit with a voltage source of $V = 24\text{ V}$ with an equivalent resistance of $R_{\text{eq}} = 12\,\Omega$. This results in a current of $I = 2\text{ A}$ from the voltage source.

Notice that resistors R_3 and R_4 are in series. They can be combined into a single equivalent resistance. One method of keeping track of the process is to include the resistors as subscripts. Here the equivalent resistance of R_3 and R_4 is

Equation:

$$R_{34} = R_3 + R_4 = 6\,\Omega + 4\,\Omega = 10\,\Omega.$$

The circuit now reduces to three resistors, shown in [link](c). Redrawing, we now see that resistors R_2 and R_{34} constitute a parallel circuit. Those two resistors can be reduced to an equivalent resistance:

Equation:

$$R_{234} = \left(\frac{1}{R_2} + \frac{1}{R_{34}}\right)^{-1} = \left(\frac{1}{10\,\Omega} + \frac{1}{10\,\Omega}\right)^{-1} = 5\,\Omega.$$

This step of the process reduces the circuit to two resistors, shown in in [link](d). Here, the circuit reduces to two resistors, which in this case are in series. These two resistors can be reduced to an equivalent resistance, which is the equivalent resistance of the circuit:

Equation:

$$R_{\text{eq}} = R_{1234} = R_1 + R_{234} = 7\,\Omega + 5\,\Omega = 12\,\Omega.$$

The main goal of this circuit analysis is reached, and the circuit is now reduced to a single resistor and single voltage source.

Now we can analyze the circuit. The current provided by the voltage source is $I = \frac{V}{R_{\text{eq}}} = \frac{24\text{ V}}{12\,\Omega} = 2\text{ A}$. This current runs through resistor R_1 and is designated as I_1. The potential drop across R_1 can be found using Ohm’s law:

Equation:

$$V_1 = I_1 R_1 = (2\text{ A})\,(7\,\Omega) = 14\text{ V}.$$

Looking at [link](c), this leaves $24\text{ V} - 14\text{ V} = 10\text{ V}$ to be dropped across the parallel combination of R_2 and R_{34}. The current through R_2 can be found using Ohm’s law:

Equation:

$$I_2 = \frac{V_2}{R_2} = \frac{10\text{ V}}{10\,\Omega} = 1\text{ A}.$$

The resistors R_3 and R_4 are in series so the currents I_3 and I_4 are equal to

Equation:

$$I_3 = I_4 = I - I_2 = 2\text{ A} - 1\text{ A} = 1\text{ A}.$$

Using Ohm’s law, we can find the potential drop across the last two resistors. The potential drops are $V_3 = I_3 R_3 = 6\text{ V}$ and $V_4 = I_4 R_4 = 4\text{ V}$. The final analysis is to look at the power supplied by the voltage source and the power dissipated by the resistors. The power dissipated by the resistors is

Equation:

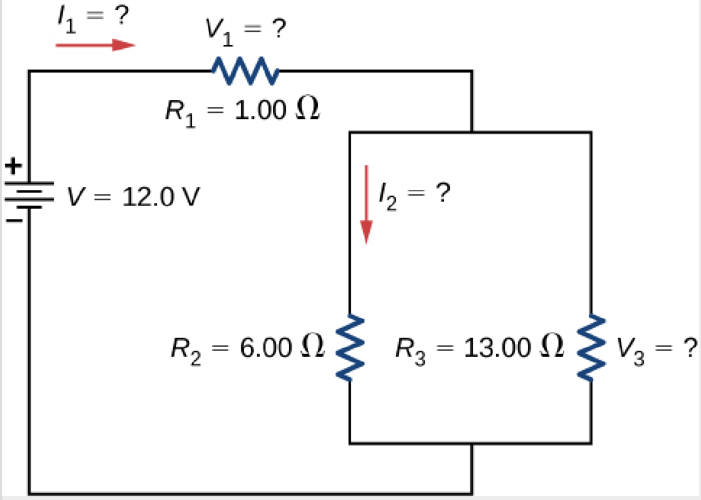
$$\begin{aligned}
P_1 &= I_1^2 R_1 = (2\text{ A})^2\,(7\,\Omega) = 28\text{ W},\\
P_2 &= I_2^2 R_2 = (1\text{ A})^2\,(10\,\Omega) = 10\text{ W},\\
P_3 &= I_3^2 R_3 = (1\text{ A})^2\,(6\,\Omega) = 6\text{ W},\\
P_4 &= I_4^2 R_4 = (1\text{ A})^2\,(4\,\Omega) = 4\text{ W},\\
P_{\text{dissipated}} &= P_1 + P_2 + P_3 + P_4 = 48\text{ W}.
\end{aligned}$$

The total energy is constant in any process. Therefore, the power supplied by the voltage source is $P_s = IV = (2\text{ A})(24\text{ V}) = 48\text{ W}$. Analyzing the power supplied to the circuit and the power dissipated by the resistors is a good check for the validity of the analysis; they should be equal.

Example:

Combining Series and Parallel Circuits

[link] shows resistors wired in a combination of series and parallel. We can consider R_1 to be the resistance of wires leading to R_2 and R_3. (a) Find the equivalent resistance of the circuit. (b) What is the potential drop V_1 across resistor R_1? (c) Find the current I_2 through resistor R_2. (d) What power is dissipated by R_2?



These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

Strategy

(a) To find the equivalent resistance, first find the equivalent resistance of the parallel connection of R_2 and R_3. Then use this result to find the equivalent resistance of the series connection with R_1.

(b) The current through R_1 can be found using Ohm's law and the voltage applied. The current through R_1 is equal to the current from the battery. The potential drop V_1 across the resistor R_1 (which represents the resistance in the connecting wires) can be found using Ohm's law.

(c) The current through R_2 can be found using Ohm's law $I_2 = \frac{V_2}{R_2}$. The voltage across R_2 can be found using $V_2 = V - V_1$.

(d) Using Ohm's law $(V_2 = I_2 R_2)$, the power dissipated by the resistor can also be found using $P_2 = I_2^2 R_2 = \frac{V_2^2}{R_2}$.

Solution

a. To find the equivalent resistance of the circuit, notice that the parallel connection of R_2 and R_3 is in series with R_1, so the equivalent resistance is
   Equation:

$$R_{\text{eq}} = R_1 + \left(\frac{1}{R_2} + \frac{1}{R_3}\right)^{-1} = 1.00\,\Omega + \left(\frac{1}{6.00\,\Omega} + \frac{1}{13.00\,\Omega}\right)^{-1} = 5.10\,\Omega.$$

The total resistance of this combination is intermediate between the pure series and pure parallel values ($20.0\,\Omega$ and $0.804\,\Omega$, respectively).

b. The current through R_1 is equal to the current supplied by the battery:
Equation:

$$I_1 = I = \frac{V}{R_{\text{eq}}} = \frac{12.0\text{ V}}{5.10\,\Omega} = 2.35\text{ A}.$$

The voltage across R_1 is
Equation:

$$V_1 = I_1 R_1 = (2.35\text{ A})(1\,\Omega) = 2.35\text{ V}.$$

The voltage applied to R_2 and R_3 is less than the voltage supplied by the battery by an amount V_1. When wire resistance is large, it can significantly affect the operation of the devices represented by R_2 and R_3.

c. To find the current through R_2, we must first find the voltage applied to it. The voltage across the two resistors in parallel is the same:
Equation:

$$V_2 = V_3 = V - V_1 = 12.0\text{ V} - 2.35\text{ V} = 9.65\text{ V}.$$

Now we can find the current I_2 through resistance R_2 using Ohm's law:
Equation:

$$I_2 = \frac{V_2}{R_2} = \frac{9.65\text{ V}}{6.00\,\Omega} = 1.61\text{ A}.$$

The current is less than the 2.00 A that flowed through R_2 when it was connected in parallel to the battery in the previous parallel circuit example.

d. The power dissipated by R_2 is given by
Equation:

$$P_2 = I_2^2 R_2 = (1.61\text{ A})^2(6.00\,\Omega) = 15.5\text{ W}.$$

Significance
The analysis of complex circuits can often be simplified by reducing the circuit to a voltage source and an equivalent resistance. Even if the entire circuit cannot be reduced to a single voltage source and a single equivalent resistance, portions of the circuit may be reduced, greatly simplifying the analysis.

Note:
Exercise:

Problem:

Check Your Understanding Consider the electrical circuits in your home. Give at least two examples of circuits that must use a combination of series and parallel circuits to operate efficiently.

Solution:

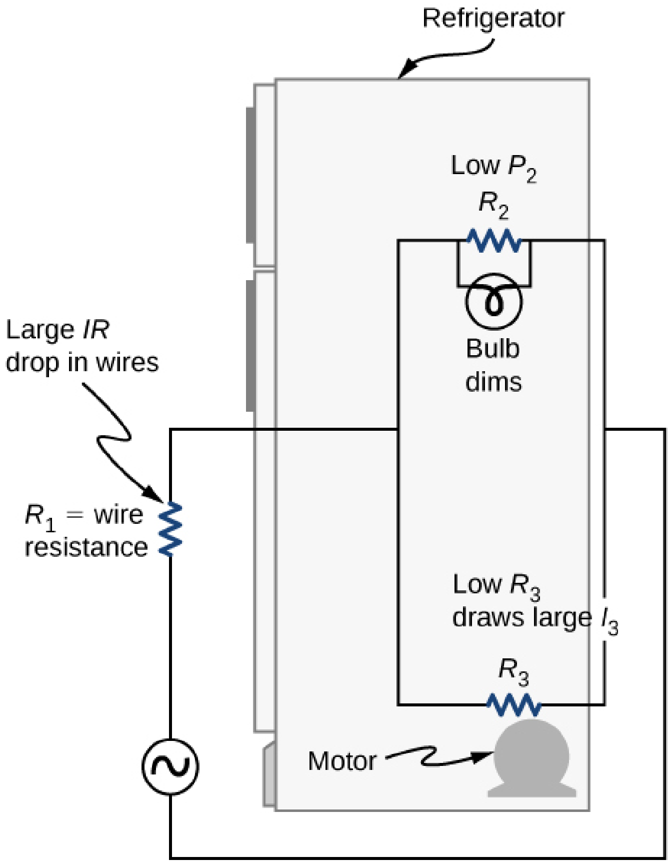
1. All the overhead lighting circuits are in parallel and connected to the main supply line, so when one bulb burns out, all the overhead lighting does not go dark. Each overhead light will have at least one switch in series with the light, so you can turn it on and off. 2. A refrigerator has a compressor and a light that goes on when the door opens. There is usually only one cord for the refrigerator to plug into the wall. The circuit containing the compressor and the circuit containing the lighting circuit are in parallel, but there is a switch in series with the light. A thermostat controls a switch that is in series with the compressor to control the temperature of the refrigerator.

Practical Implications

One implication of this last example is that resistance in wires reduces the current and power delivered to a resistor. If wire resistance is relatively large, as in a worn (or a very long) extension cord, then this loss can be significant. If a large current is drawn, the *IR* drop in the wires can also be significant and may become apparent from the heat generated in the cord.

For example, when you are rummaging in the refrigerator and the motor comes on, the refrigerator light dims momentarily. Similarly, you can see the passenger compartment light dim when you start the engine of your car (although this may be due to resistance inside the battery itself).

What is happening in these high-current situations is illustrated in [link]. The device represented by R_3 has a very low resistance, so when it is switched on, a large current flows. This increased current causes a larger *IR* drop in the wires represented by R_1, reducing the voltage across the light bulb (which is R_2), which then dims noticeably.



Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant *IR* drop in the wires and reduces the voltage across the light.

Note:

Problem-Solving Strategy: Series and Parallel Resistors

1. Draw a clear circuit diagram, labeling all resistors and voltage sources. This step includes a list of the known values for the problem, since they are labeled in your circuit diagram.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
3. Determine whether resistors are in series, parallel, or a combination of both series and parallel. Examine the circuit diagram to make this assessment. Resistors are in series if the same current must pass sequentially through them.
4. Use the appropriate list of major features for series or parallel connections to solve for the unknowns. There is one list for series and another for parallel.
5. Check to see whether the answers are reasonable and consistent.

Example:

Combining Series and Parallel Circuits

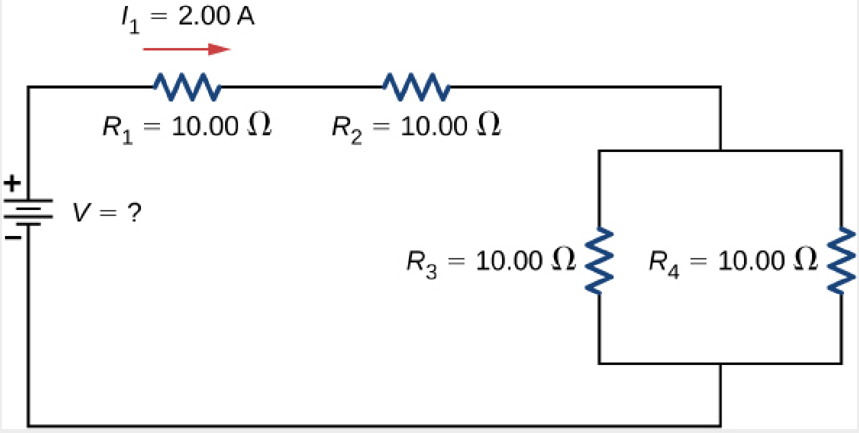
Two resistors connected in series (R_1, R_2) are connected to two resistors that are connected in parallel (R_3, R_4). The series-parallel combination is connected to a battery. Each resistor has a resistance of 10.00 Ohms. The wires connecting the resistors and battery have negligible resistance. A current of 2.00 Amps runs through resistor R_1. What is the voltage supplied by the voltage source?

Strategy

Use the steps in the preceding problem-solving strategy to find the solution for this example.

Solution

1. Draw a clear circuit diagram ([link]).



To find the unknown voltage, we must first find the equivalent resistance of the circuit.

2. The unknown is the voltage of the battery. In order to find the voltage supplied by the battery, the equivalent resistance must be found.
3. In this circuit, we already know that the resistors R_1 and R_2 are in series and the resistors R_3 and R_4 are in parallel. The equivalent resistance of the parallel configuration of the resistors R_3 and R_4 is in series with the series configuration of resistors R_1 and R_2.
4. The voltage supplied by the battery can be found by multiplying the current from the battery and the equivalent resistance of the circuit. The current from the battery is equal to the current through R_1 and is equal to 2.00 A. We need to find the equivalent resistance by reducing the circuit. To reduce the circuit, first consider the two resistors in parallel. The equivalent resistance is $R_{34} = \left(\frac{1}{10.00\,\Omega} + \frac{1}{10.00\,\Omega}\right)^{-1} = 5.00\,\Omega$. This parallel combination is in series with the other two resistors, so the equivalent resistance of the circuit is $R_{\text{eq}} = R_1 + R_2 + R_{34} = 25.00\,\Omega$. The voltage supplied by the battery is therefore $V = IR_{\text{eq}} = 2.00\text{ A}\,(25.00\,\Omega) = 50.00\text{ V}$.
5. One way to check the consistency of your results is to calculate the power supplied by the battery and the power dissipated by the resistors. The power supplied by the battery is $P_{\text{batt}} = IV = 100.00\text{ W}$.
   Since they are in series, the current through R_2 equals the current through R_1. Since $R_3 = R_4$, the current through each will be 1.00 Amps. The power dissipated by the resistors is equal to the sum of the power dissipated by each resistor:
   Equation:

$$P = I_1^2 R_1 + I_2^2 R_2 + I_3^2 R_3 + I_4^2 R_4 = 40.00\text{ W} + 40.00\text{ W} + 10.00\text{ W} + 10.00\text{ W} = 100.00\text{ W}.$$

Since the power dissipated by the resistors equals the power supplied by the battery, our solution seems consistent.

Significance
If a problem has a combination of series and parallel, as in this example, it can be reduced in steps by using the preceding problem-solving strategy and by considering individual groups of series or parallel connections. When finding R_{eq} for a parallel connection, the reciprocal must be taken with care. In addition, units and numerical results must be reasonable. Equivalent series resistance should be greater, whereas equivalent parallel resistance should be smaller, for example. Power should be greater for the same devices in parallel compared with series, and so on.

Summary

- The equivalent resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances: $R_{\text{s}} = R_1 + R_2 + R_3 + \cdots = \sum_{i=1}^{N} R_i.$
- Each resistor in a series circuit has the same amount of current flowing through it.
- The potential drop, or power dissipation, across each individual resistor in a series is different, and their combined total is the power source input.
- The equivalent resistance of an electrical circuit with resistors wired in parallel is less than the lowest resistance of any of the components and can be determined using the formula
  Equation:

$$R_{\text{eq}} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots\right)^{-1} = \left(\sum_{i=1}^{N} \frac{1}{R_i}\right)^{-1}.$$

- Each resistor in a parallel circuit has the same full voltage of the source applied to it.
- The current flowing through each resistor in a parallel circuit is different, depending on the resistance.
- If a more complex connection of resistors is a combination of series and parallel, it can be reduced to a single equivalent resistance by identifying its various parts as series or parallel, reducing each to its equivalent, and continuing until a single resistance is eventually reached.

Conceptual Questions

Exercise:

Problem: A voltage occurs across an open switch. What is the power dissipated by the open switch?

Exercise:

Problem:

The severity of a shock depends on the magnitude of the current through your body. Would you prefer to be in series or in parallel with a resistance, such as the heating element of a toaster, if you were shocked by it? Explain.

Solution:

It would probably be better to be in series because the current will be less than if it were in parallel.

Exercise:

Problem:

Suppose you are doing a physics lab that asks you to put a resistor into a circuit, but all the resistors supplied have a larger resistance than the requested value. How would you connect the available resistances to attempt to get the smaller value asked for?

Exercise:

Problem:

Some light bulbs have three power settings (not including zero), obtained from multiple filaments that are individually switched and wired in parallel. What is the minimum number of filaments needed for three power settings?

Solution:

two filaments, a low resistance and a high resistance, connected in parallel

Problems

Exercise:

Problem:

(a) What is the resistance of a $1.00 \times 10^{2}\text{-}\Omega$, a $2.50\text{-k}\Omega$, and a $4.00\text{-k}\Omega$ resistor connected in series? (b) In parallel?

Exercise:

Problem:

What are the largest and smallest resistances you can obtain by connecting a $36.0\text{-}\Omega$, a $50.0\text{-}\Omega$, and a $700\text{-}\Omega$ resistor together?

Solution:

largest, $786\ \Omega$, smallest, $20.32\ \Omega$

Exercise:

Problem:

An 1800-W toaster, a 1400-W speaker, and a 75-W lamp are plugged into the same outlet in a 15-A fuse and 120-V circuit. (The three devices are in parallel when plugged into the same socket.) (a) What current is drawn by each device? (b) Will this combination blow the 15-A fuse?

Exercise:

Problem:

Your car’s 30.0-W headlight and 2.40-kW starter are ordinarily connected in parallel in a 12.0-V system. What power would one headlight and the starter consume if connected in series to a 12.0-V battery? (Neglect any other resistance in the circuit and any change in resistance in the two devices.)

Solution:

29.6 W

Exercise:

Problem:

(a) Given a 48.0-V battery and $24.0\text{-}\Omega$ and $96.0\text{-}\Omega$ resistors, find the current and power for each when connected in series. (b) Repeat when the resistances are in parallel.

Exercise:

Problem:

Referring to the example combining series and parallel circuits and [link], calculate I_3 in the following two different ways: (a) from the known values of I and I_2; (b) using Ohm's law for R_3. In both parts, explicitly show how you follow the steps in the [link].

Solution:

a. 0.74 A; b. 0.742 A

Exercise:

Problem:

Referring to [link], (a) Calculate P_3 and note how it compares with P_3 found in the first two example problems in this module. (b) Find the total power supplied by the source and compare it with the sum of the powers dissipated by the resistors.

Exercise:

Problem:

Refer to [link] and the discussion of lights dimming when a heavy appliance comes on. (a) Given the voltage source is 120 V, the wire resistance is $0.800\ \Omega$, and the bulb is nominally 75.0 W, what power will the bulb dissipate if a total of 15.0 A passes through the wires when the motor comes on? Assume negligible change in bulb resistance. (b) What power is consumed by the motor?

Solution:

a. 60.8 W; b. 1.56 kW

Exercise:

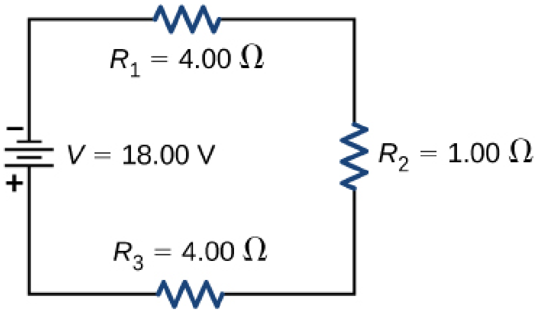
Problem:

Show that if two resistors R_1 and R_2 are combined and one is much greater than the other $(R_1 \gg R_2)$, (a) their series resistance is very nearly equal to the greater resistance R_1 and (b) their parallel resistance is very nearly equal to the smaller resistance R_2.

Exercise:

Problem:

Consider the circuit shown below. The terminal voltage of the battery is $V = 18.00\ \text{V}$. (a) Find the equivalent resistance of the circuit. (b) Find the current through each resistor. (c) Find the potential drop across each resistor. (d) Find the power dissipated by each resistor. (e) Find the power supplied by the battery.



Solution:

a. $R_s = 9.00\ \Omega$; b.$I_1 = I_2 = I_3 = 2.00\ \text{A}$;
c. $V_1 = 8.00\ \text{V}, V_2 = 2.00\ \text{V}, V_3 = 8.00\ \text{V}$; d. $P_1 = 16.00\ \text{W}, P_2 = 4.00\ \text{W}, P_3 = 16.00\ \text{W}$; e. $P = 36.00\ \text{W}$

Glossary

equivalent resistance	resistance of a combination of resistors; it can be thought of as the resistance of a single resistor that can replace a combination of resistors in a series and/or parallel circuit

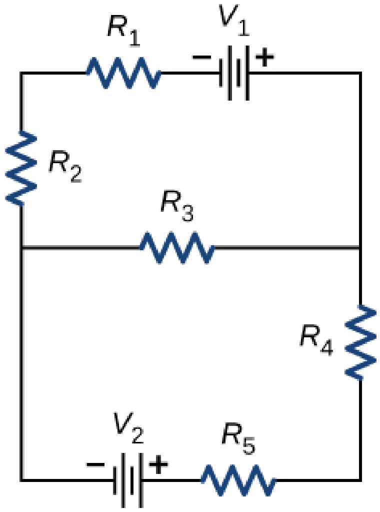
Kirchhoff's Rules

By the end of the section, you will be able to:

- State Kirchhoff’s junction rule
- State Kirchhoff’s loop rule
- Analyze complex circuits using Kirchhoff’s rules

We have just seen that some circuits may be analyzed by reducing a circuit to a single voltage source and an equivalent resistance. Many complex circuits cannot be analyzed with the series-parallel techniques developed in the preceding sections. In this section, we elaborate on the use of Kirchhoff’s rules to analyze more complex circuits. For example, the circuit in [link] is known as a multi-loop circuit, which consists of junctions. A junction, also known as a node, is a connection of three or more wires. In this circuit, the previous methods cannot be used, because not all the resistors are in clear series or parallel configurations that can be reduced. Give it a try. The resistors R_1 and R_2 are in series and can be reduced to an equivalent resistance. The same is true of resistors R_4 and R_5. But what do you do then?

Even though this circuit cannot be analyzed using the methods already learned, two circuit analysis rules can be used to analyze any circuit, simple or complex. The rules are known as **Kirchhoff’s rules**, after their inventor Gustav Kirchhoff (1824–1887).



This circuit cannot be reduced to a combination of series and parallel connections. However, we can use Kirchhoff’s rules to analyze it.

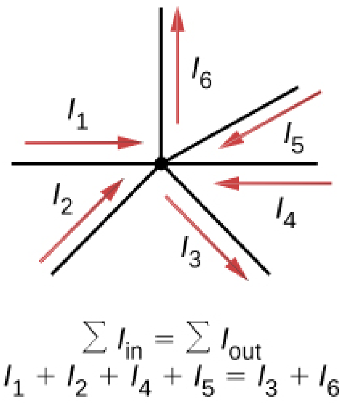
Note:
Kirchhoff’s Rules

- Kirchhoff’s first rule—the junction rule. The sum of all currents entering a junction must equal the sum of all currents leaving the junction:
  Equation:
  $$\sum I_{\text{in}} = \sum I_{\text{out}}.$$
- Kirchhoff’s second rule—the loop rule. The algebraic sum of changes in potential around any closed circuit path (loop) must be zero:
  Equation:
  $$\sum V = 0.$$

We now provide explanations of these two rules, followed by problem-solving hints for applying them and a worked example that uses them.

Kirchhoff’s First Rule

Kirchhoff’s first rule (the **junction rule**) applies to the charge entering and leaving a junction ([link]). As stated earlier, a junction, or node, is a connection of three or more wires. Current is the flow of charge, and charge is conserved; thus, whatever charge flows into the junction must flow out.

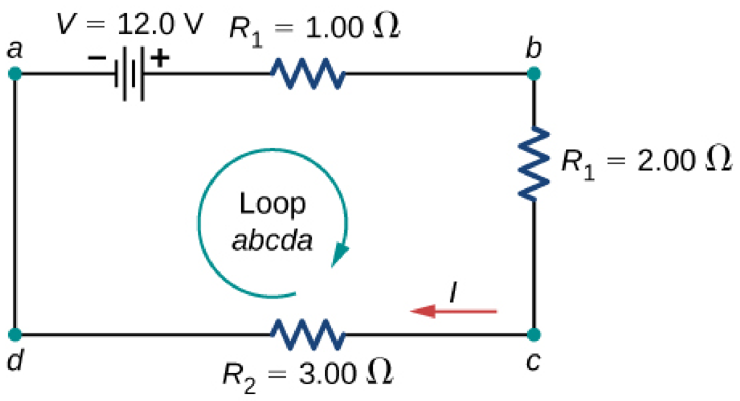

I_6
I_1
I_5
I_2
I_4
I_3
$\sum I_{\text{in}} = \sum I_{\text{out}}$
$I_1 + I_2 + I_4 + I_5 = I_3 + I_6$

Charge must be conserved, so the sum of currents into a junction must be equal to the sum of currents out of the junction.

Although it is an over-simplification, an analogy can be made with water pipes connected in a plumbing junction. If the wires in [link] were replaced by water pipes, and the water was assumed to be incompressible, the volume of water flowing into the junction must equal the volume of water flowing out of the junction.

Kirchhoff's Second Rule

Kirchhoff's second rule (the **loop rule**) applies to potential differences. The loop rule is stated in terms of potential V rather than potential energy, but the two are related since $U=qV$. In a closed loop, whatever energy is supplied by a voltage source, the energy must be transferred into other forms by the devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. Kirchhoff's loop rule states that the algebraic sum of potential differences, including voltage supplied by the voltage sources and resistive elements, in any loop must be equal to zero. For example, consider a simple loop with no junctions, as in [link].

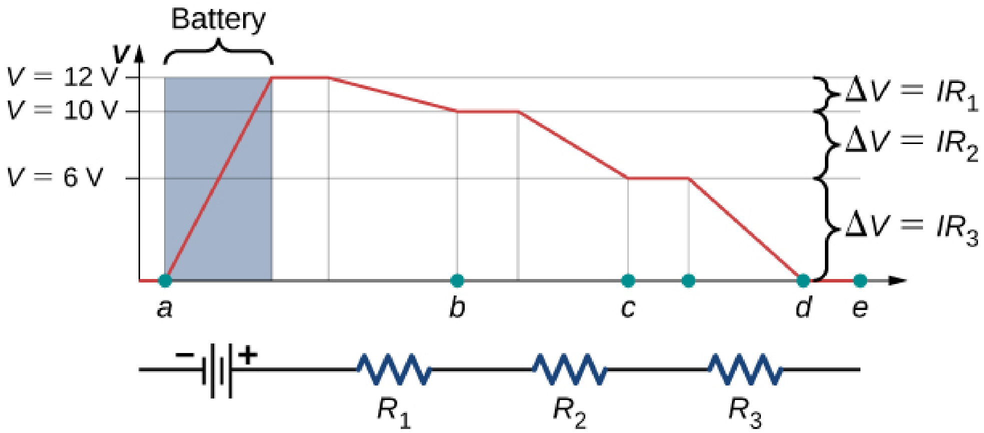


A simple loop with no junctions. Kirchhoff's loop rule states that the algebraic sum of the voltage differences is equal to zero.

The circuit consists of a voltage source and three external load resistors. The labels *a*, *b*, *c*, and *d* serve as references, and have no other significance. The usefulness of these labels will become apparent soon. The loop is designated as Loop *abcda*, and the labels help keep track of the voltage differences as we travel around the circuit. Start at point *a* and travel to point *b*.

The voltage of the voltage source is added to the equation and the potential drop of the resistor R_1 is subtracted. From point *b* to *c*, the potential drop across R_2 is subtracted. From *c* to *d*, the potential drop across R_3 is subtracted. From points *d* to *a*, nothing is done because there are no components.

[link] shows a graph of the voltage as we travel around the loop. Voltage increases as we cross the battery, whereas voltage decreases as we travel across a resistor. The potential drop, or change in the electric potential, is equal to the current through the resistor times the resistance of the resistor. Since the wires have negligible resistance, the voltage remains constant as we cross the wires connecting the components.



A voltage graph as we travel around the circuit. The voltage increases as we cross the battery and decreases as we cross each resistor. Since the resistance of the wire is quite small, we assume that the voltage remains constant as we cross the wires connecting the components.

Then Kirchhoff’s loop rule states
Equation:

$$V - IR_1 - IR_2 - IR_3 = 0.$$

The loop equation can be used to find the current through the loop:
Equation:

$$I = \frac{V}{R_1 + R_2 + R_2} = \frac{12.00\ \text{V}}{1.00\ \Omega + 2.00\ \Omega + 3.00\ \Omega} = 2.00\ \text{A}.$$

This loop could have been analyzed using the previous methods, but we will demonstrate the power of Kirchhoff’s method in the next section.

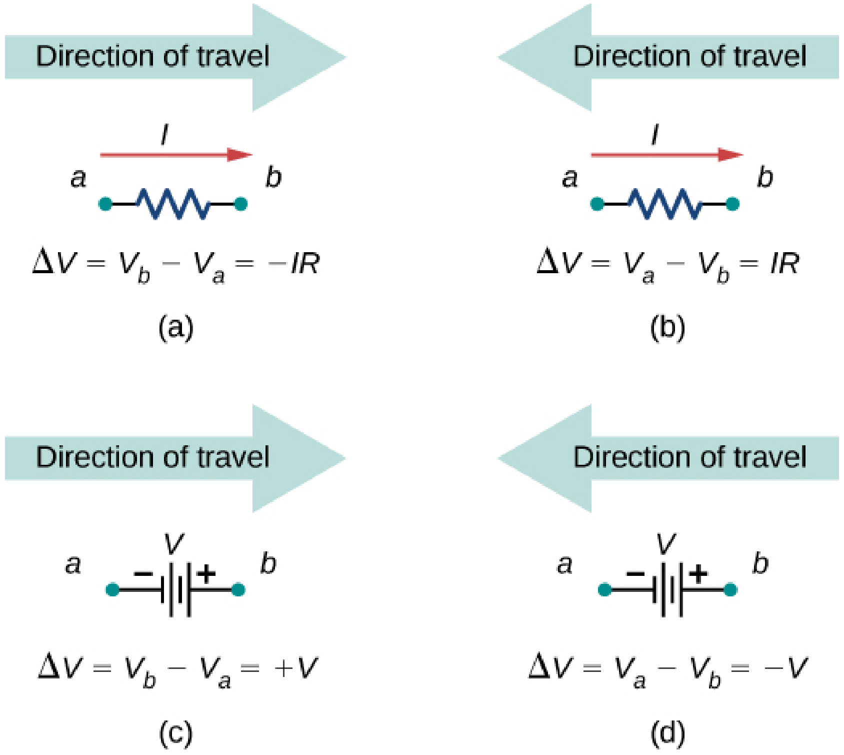
Applying Kirchhoff’s Rules

By applying Kirchhoff’s rules, we generate a set of linear equations that allow us to find the unknown values in circuits. These may be currents, voltages, or resistances. Each time a rule is applied, it produces an equation. If there are as many independent equations as unknowns, then the problem can be solved.

Using Kirchhoff’s method of analysis requires several steps, as listed in the following procedure.

Note:
Problem-Solving Strategy: Kirchhoff’s Rules

1. Label points in the circuit diagram using lowercase letters *a*, *b*, *c*, …. These labels simply help with orientation.
2. Locate the junctions in the circuit. The junctions are points where three or more wires connect. Label each junction with the currents and directions into and out of it. Make sure at least one current points into the junction and at least one current points out of the junction.
3. Choose the loops in the circuit. Every component must be contained in at least one loop, but a component may be contained in more than one loop.
4. Apply the junction rule. Again, some junctions should not be included in the analysis. You need only use enough nodes to include every current.
5. Apply the loop rule. Use the map in [link].

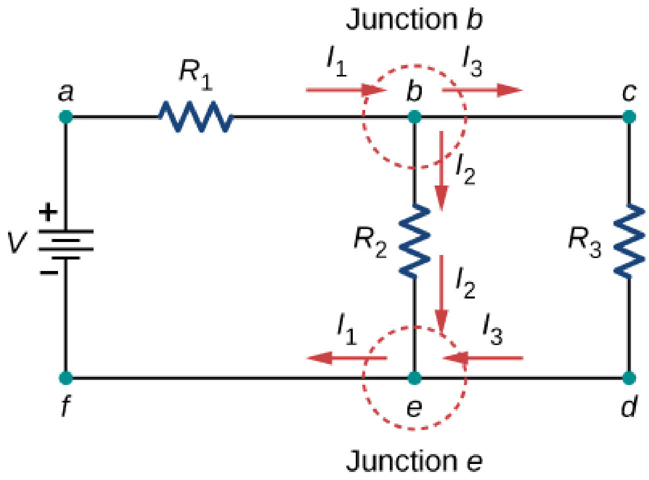


Each of these resistors and voltage sources is traversed from *a* to *b*. (a) When moving across a resistor in the same direction as the current flow, subtract the potential drop. (b) When moving across a resistor in the opposite direction as the current flow, add the potential drop. (c) When moving across a voltage source from the negative terminal to the positive terminal, add the potential drop. (d) When moving across a voltage source from the positive terminal to the negative terminal, subtract the potential drop.

Let’s examine some steps in this procedure more closely. When locating the junctions in the circuit, do not be concerned about the direction of the currents. If the direction of current flow is not obvious, choosing any direction is sufficient as long as at least one current points into the junction and at least one current points out of the junction. If the arrow is in the opposite direction of the conventional current flow, the result for the current in question will be negative but the answer will still be correct.

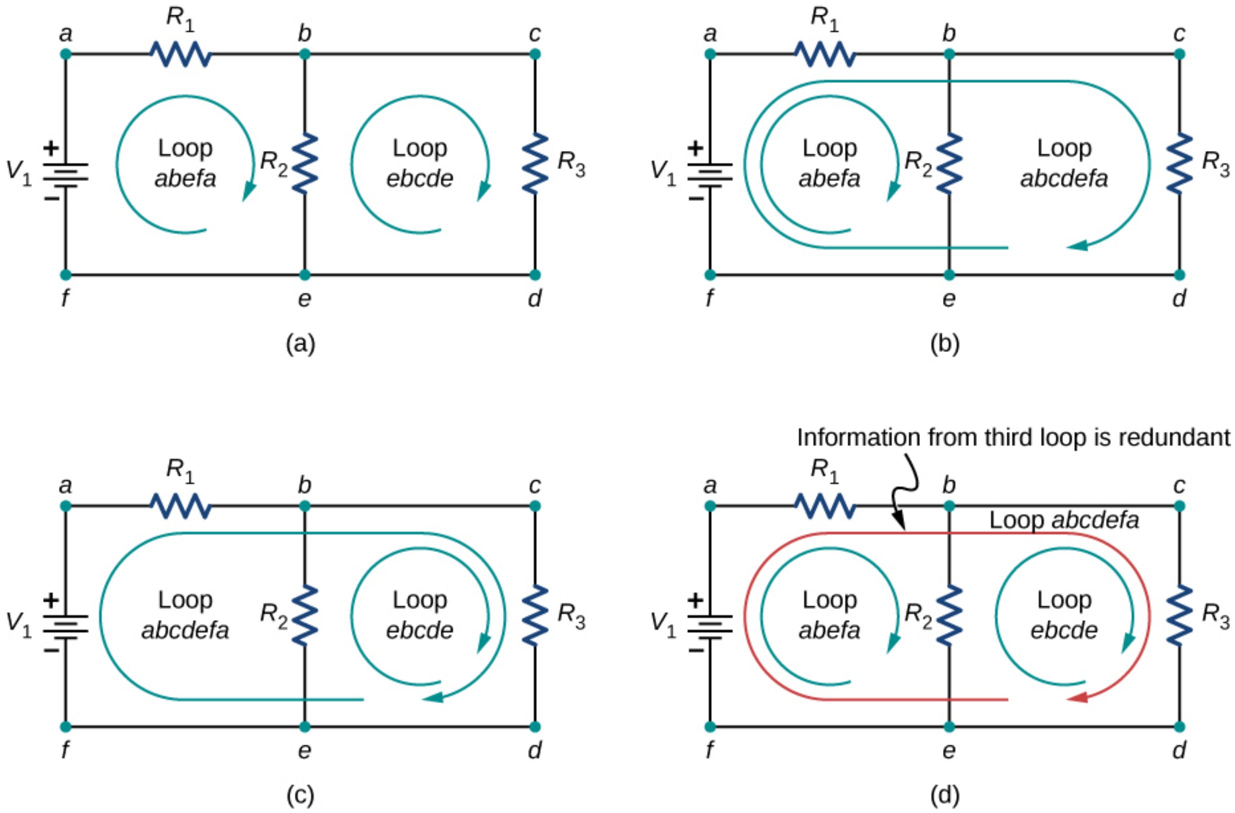
The number of nodes depends on the circuit. Each current should be included in a node and thus included in at least one junction equation. Do not include nodes that are not linearly independent, meaning nodes that contain the same information.

Consider [link]. There are two junctions in this circuit: Junction *b* and Junction *e*. Points *a*, *c*, *d*, and *f* are not junctions, because a junction must have three or more connections. The equation for Junction *b* is $I_1 = I_2 + I_3$, and the equation for Junction *e* is $I_2 + I_3 = I_1$. These are equivalent equations, so it is necessary to keep only one of them.



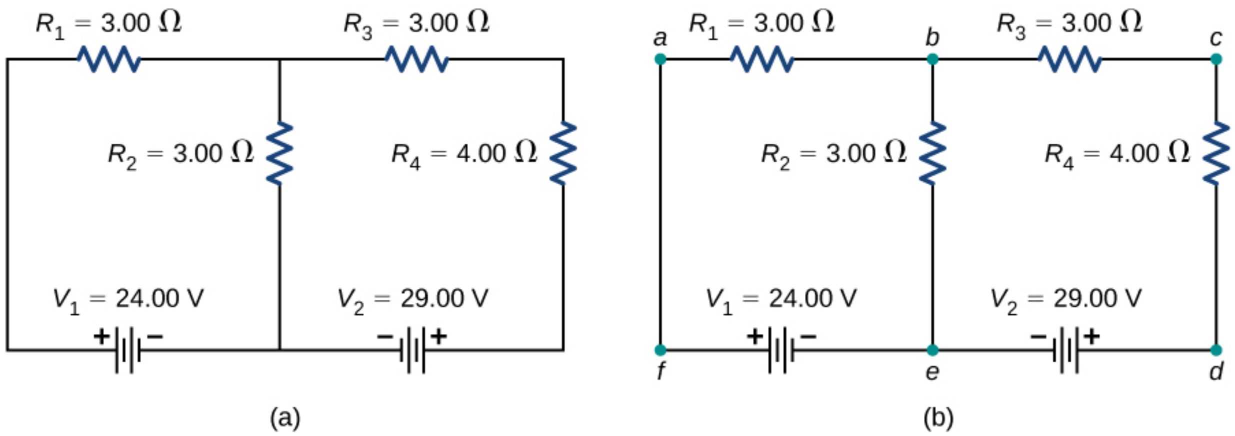
At first glance, this circuit contains two junctions, Junction *b* and Junction *e*, but only one should be considered because their junction equations are equivalent.

When choosing the loops in the circuit, you need enough loops so that each component is covered once, without repeating loops. [link] shows four choices for loops to solve a sample circuit; choices (a), (b), and (c) have a sufficient amount of loops to solve the circuit completely. Option (d) reflects more loops than necessary to solve the circuit.


a
b
c
d
e
f
R_1
R_2
R_3
V_1
+
–
Loop abefa
Loop ebcde
Loop abcdefa
Information from third loop is redundant
(a)
(b)
(c)
(d)

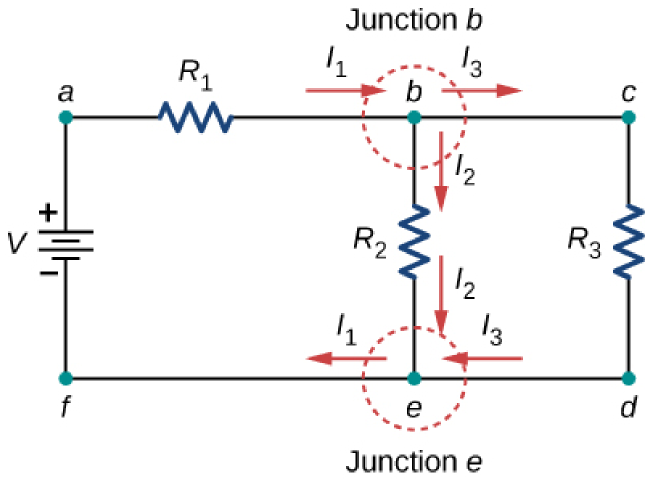
Panels (a)–(c) are sufficient for the analysis of the circuit. In each case, the two loops shown contain all the circuit elements necessary to solve the circuit completely. Panel (d) shows three loops used, which is more than necessary. Any two loops in the system will contain all information needed to solve the circuit. Adding the third loop provides redundant information.

Consider the circuit in [link](a). Let us analyze this circuit to find the current through each resistor. First, label the circuit as shown in part (b).


R1 = 3.00 Ω
R3 = 3.00 Ω
R2 = 3.00 Ω
R4 = 4.00 Ω
V1 = 24.00 V
V2 = 29.00 V
+
−
−
+
(a)
a
b
c
R1 = 3.00 Ω
R3 = 3.00 Ω
R2 = 3.00 Ω
R4 = 4.00 Ω
V1 = 24.00 V
V2 = 29.00 V
+
−
−
+
f
e
d
(b)

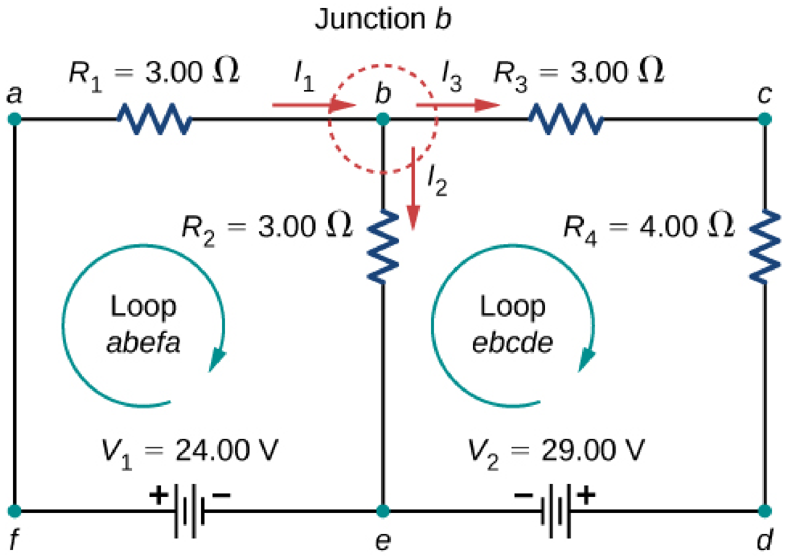
(a) A multi-loop circuit. (b) Label the circuit to help with orientation.

Next, determine the junctions. In this circuit, points *b* and *e* each have three wires connected, making them junctions. Start to apply Kirchhoff’s junction rule $\left(\sum I_{\text{in}} = \sum I_{\text{out}}\right)$ by drawing arrows representing the currents and labeling each arrow, as shown in [link](b). Junction *b* shows that $I_1 = I_2 + I_3$ and Junction *e* shows that $I_2 + I_3 = I_1$. Since Junction *e* gives the same information of Junction *b*, it can be disregarded. This circuit has three unknowns, so we need three linearly independent equations to analyze it.


Junction b
I1
I3
a
R1
b
c
I2
V
+
−
R2
R3
I2
I1
I3
f
e
d
Junction e

(a) This circuit has two junctions, labeled *b* and *e*, but only node *b* is used in the analysis. (b) Labeled arrows represent the currents into and out of the junctions.

Next we need to choose the loops. In [link], Loop *abefa* includes the voltage source V_1 and resistors R_1 and R_2. The loop starts at point *a*, then travels through points *b*, *e*, and *f*, and then back to point *a*. The second loop, Loop *ebcde*, starts at point *e* and includes resistors R_2 and R_3, and the voltage source V_2.



Choose the loops in the circuit.

Now we can apply Kirchhoff's loop rule, using the map in [link]. Starting at point *a* and moving to point *b*, the resistor R_1 is crossed in the same direction as the current flow I_1, so the potential drop I_1R_1 is subtracted. Moving from point *b* to point *e*, the resistor R_2 is crossed in the same direction as the current flow I_2 so the potential drop I_2R_2 is subtracted. Moving from point *e* to point *f*, the voltage source V_1 is crossed from the negative terminal to the positive terminal, so V_1 is added. There are no components between points *f* and *a*. The sum of the voltage differences must equal zero:

Equation:

$$\text{Loop } abefa: \quad -I_1R_1 - I_2R_2 + V_1 = 0 \text{ or } V_1 = I_1R_1 + I_2R_2.$$

Finally, we check loop *ebcde*. We start at point *e* and move to point *b*, crossing R_2 in the opposite direction as the current flow I_2. The potential drop I_2R_2 is added. Next, we cross R_3 and R_4 in the same direction as the current flow I_3 and subtract the potential drops I_3R_3 and I_3R_4. Note that the current is the same through resistors R_3 and R_4, because they are connected in series. Finally, the voltage source is crossed from the positive terminal to the negative terminal, and the voltage source V_2 is subtracted. The sum of these voltage differences equals zero and yields the loop equation

Equation:

$$\text{Loop } ebcde: I_2 R_2 - I_3 \left(R_3 + R_4\right) - V_2 = 0.$$

We now have three equations, which we can solve for the three unknowns.
Equation:

$$\begin{aligned}&(1)\,\text{Junction } b: I_1 - I_2 - I_3 = 0.\\&(2)\,\text{Loop } abefa: \quad I_1 R_1 + I_2 R_2 = V_1.\\&(3)\,\text{Loop } ebcde: I_2 R_2 - I_3 \left(R_3 + R_4\right) = V_2.\end{aligned}$$

To solve the three equations for the three unknown currents, start by eliminating current I_2. First add Eq. (1) times R_2 to Eq. (2). The result is labeled as Eq. (4):
Equation:

$$(R_1 + R_2) I_1 - R_2 I_3 = V_1.$$

$$(4)\; 6\,\Omega I_1 - 3\,\Omega I_3 = 24\,\text{V}.$$

Next, subtract Eq. (3) from Eq. (2). The result is labeled as Eq. (5):
Equation:

$$I_1 R_1 + I_3 \left(R_3 + R_4\right) = V_1 - V_2.$$

$$(5)\; 3\,\Omega I_1 + 7\,\Omega I_3 = -5\,\text{V}.$$

We can solve Eqs. (4) and (5) for current I_1. Adding seven times Eq. (4) and three times Eq. (5) results in $51\,\Omega I_1 = 153\,\text{V}$, or $I_1 = 3.00\,\text{A}$. Using Eq. (4) results in $I_3 = -2.00\,\text{A}$. Finally, Eq. (1) yields $I_2 = I_1 - I_3 = 5.00\,\text{A}$. One way to check that the solutions are consistent is to check the power supplied by the voltage sources and the power dissipated by the resistors:
Equation:

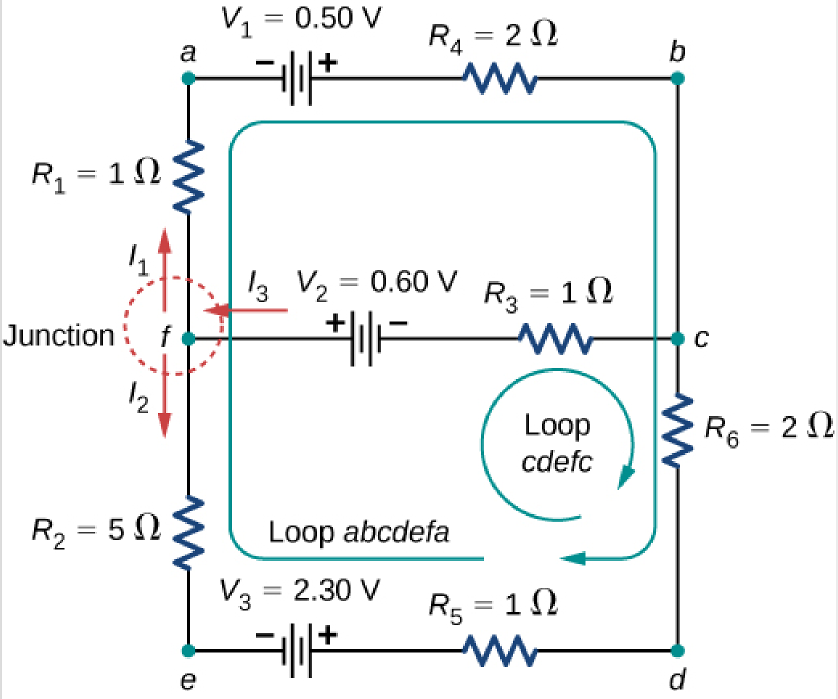
$$\begin{aligned}&P_{\text{in}} = I_1 V_1 + I_3 V_2 = 130\,\text{W},\\&P_{\text{out}} = I_1^2 R_1 + I_2^2 R_2 + I_3^2 R_3 + I_3^2 R_4 = 130\,\text{W}.\end{aligned}$$

Note that the solution for the current I_3 is negative. This is the correct answer, but suggests that the arrow originally drawn in the junction analysis is the direction opposite of conventional current flow. The power supplied by the second voltage source is 58 W and not −58 W.

Example:

Calculating Current by Using Kirchhoff's Rules

Find the currents flowing in the circuit in [link].



$V_1 = 0.50$ V
$R_4 = 2\ \Omega$
a
b
$R_1 = 1\ \Omega$
I_1
I_3
$V_2 = 0.60$ V
$R_3 = 1\ \Omega$
Junction
f
c
I_2
Loop cdefc
$R_6 = 2\ \Omega$
$R_2 = 5\ \Omega$
Loop abcdefa
$V_3 = 2.30$ V
$R_5 = 1\ \Omega$
e
d

This circuit is combination of series and parallel configurations of resistors and voltage sources. This circuit cannot be analyzed using the techniques discussed in Electromotive Force but can be analyzed using Kirchhoff's rules.

Strategy

This circuit is sufficiently complex that the currents cannot be found using Ohm's law and the series-parallel techniques—it is necessary to use Kirchhoff's rules. Currents have been labeled I_1,I_2, and I_3 in the figure, and assumptions have been made about their directions. Locations on the diagram have been labeled with letters *a* through *h*. In the solution, we apply the junction and loop rules, seeking three independent equations to allow us to solve for the three unknown currents.

Solution

Applying the junction and loop rules yields the following three equations. We have three unknowns, so three equations are required.

Equation:

$$\begin{aligned}&\text{Junction } c: \ I_1 + I_2 = I_3. \\ &\text{Loop } abcdefa: \ I_1\left(R_1 + R_4\right) - I_2\left(R_2 + R_5 + R_6\right) = V_1 - V_3. \\ &\text{Loop } cdefc: \ I_2\left(R_2 + R_5 + R_6\right) + I_3 R_3 = V_2 + V_3.\end{aligned}$$

Simplify the equations by placing the unknowns on one side of the equations.

Equation:

$$\begin{aligned}&\text{Junction } c: \ I_1 + I_2 - I_3 = 0. \\ &\text{Loop } abcdefa: \ I_1\,(3\,\Omega) - I_2\,(8\,\Omega) = 0.5\text{ V} - 2.30\text{ V}. \\ &\text{Loop } cdefc: \ I_2\,(8\,\Omega) + I_3\,(1\,\Omega) = 0.6\text{ V} + 2.30\text{ V}.\end{aligned}$$

Simplify the equations. The first loop equation can be simplified by dividing both sides by 3.00. The second loop equation can be simplified by dividing both sides by 6.00.

Equation:

$$\begin{aligned}&\text{Junction } c: \ I_1 + I_2 - I_3 = 0. \\ &\text{Loop } abcdefa: \ I_1\,(3\,\Omega) - I_2\,(8\,\Omega) = -1.8\text{ V}. \\ &\text{Loop } cdefc: \ I_2\,(8\,\Omega) + I_3\,(1\,\Omega) = 2.9\text{ V}.\end{aligned}$$

The results are

Equation:

$$I_1 = 0.20\text{ A}, \quad I_2 = 0.30\text{ A}, \quad I_3 = 0.50\text{ A}.$$

Significance

A method to check the calculations is to compute the power dissipated by the resistors and the power supplied by the voltage sources:

Equation:

$$\begin{aligned}&P_{R_1} = I_1^2 R_1 = 0.04\text{ W}. \\ &P_{R_2} = I_2^2 R_2 = 0.45\text{ W}. \\ &P_{R_3} = I_3^2 R_3 = 0.25\text{ W}. \\ &P_{R_4} = I_1^2 R_4 = 0.08\text{ W}. \\ &P_{R_5} = I_2^2 R_5 = 0.09\text{ W}. \\ &P_{R_6} = I_2^2 R_6 = 0.18\text{ W}. \\ &P_{\text{dissipated}} = 1.09\text{ W}. \\ &P_{\text{source}} = I_1 V_1 + I_2 V_3 + I_3 V_2 = 0.10\text{ W} + 0.69\text{ W} + 0.30\text{ W} = 1.09\text{ W}.\end{aligned}$$

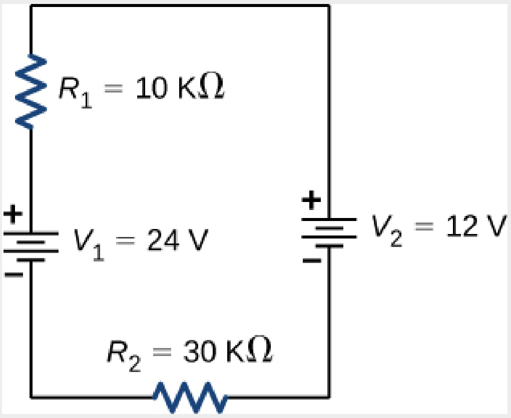
The power supplied equals the power dissipated by the resistors.

Note:

Exercise:

Problem:

Check Your Understanding In considering the following schematic and the power supplied and consumed by a circuit, will a voltage source always provide power to the circuit, or can a voltage source consume power?


$R_1 = 10\ \text{K}\Omega$
$V_1 = 24\ \text{V}$
$V_2 = 12\ \text{V}$
$R_2 = 30\ \text{K}\Omega$
+
−
+
−

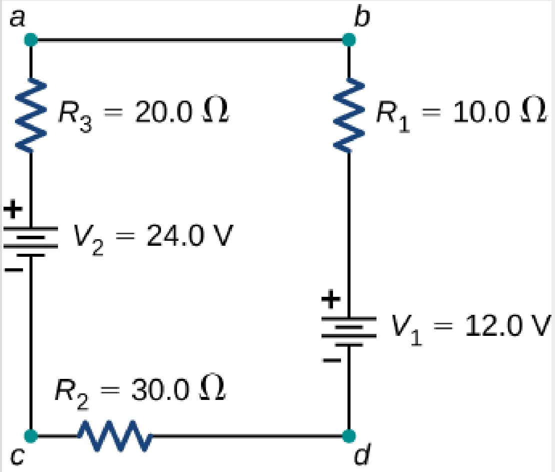
Solution:

The circuit can be analyzed using Kirchhoff's loop rule. The first voltage source supplies power: $P_{\text{in}} = IV_1 = 7.20\,\text{mW}$. The second voltage source consumes power: $P_{\text{out}} = IV_2 + I^2R_1 + I^2R_2 = 7.2\,\text{mW}$.

Example:

Calculating Current by Using Kirchhoff's Rules

Find the current flowing in the circuit in [link].


a
b
$R_3 = 20.0\ \Omega$
$R_1 = 10.0\ \Omega$
$V_2 = 24.0\ \text{V}$
$V_1 = 12.0\ \text{V}$
$R_2 = 30.0\ \Omega$
c
d
+
−
+
−

This circuit consists of three resistors and two batteries connected in series.

Note that the batteries are connected
with opposite polarities.

Strategy

This circuit can be analyzed using Kirchhoff's rules. There is only one loop and no nodes. Choose the direction of current flow. For this example, we will use the clockwise direction from point *a* to point *b*. Consider Loop *abcda* and use [link] to write the loop equation. Note that according to [link], battery V_1 will be added and battery V_2 will be subtracted.

Solution

Applying the junction rule yields the following three equations. We have one unknown, so one equation is required:

Equation:

$$\text{Loop } abcda: \; -IR_1 - V_1 - IR_2 + V_2 - IR_3 = 0.$$

Simplify the equations by placing the unknowns on one side of the equations. Use the values given in the figure.

Equation:

$$I\left(R_1 + R_2 + R_3\right) = V_2 - V_1.$$

$$I = \frac{V_2 - V_1}{R_1 + R_2 + R_3} = \frac{24\text{ V} - 12\text{ V}}{10.0\,\Omega + 30.0\,\Omega + 10.0\,\Omega} = 0.20\text{ A}.$$

Significance

The power dissipated or consumed by the circuit equals the power supplied to the circuit, but notice that the current in the battery V_1 is flowing through the battery from the positive terminal to the negative terminal and consumes power.

Equation:

$$\begin{aligned}
P_{R_1} &= I^2R_1 = 0.40\text{ W}\\
P_{R_2} &= I^2R_2 = 1.20\text{ W}\\
P_{R_3} &= I^2R_3 = 0.80\text{ W}\\
P_{V_1} &= IV_1 = 2.40\text{ W}\\
P_{\text{dissipated}} &= 4.80\text{ W}\\
P_{\text{source}} &= IV_2 = 4.80\text{ W}
\end{aligned}$$

The power supplied equals the power dissipated by the resistors and consumed by the battery V_1.

Note:

Exercise:

Problem:

Check Your Understanding When using Kirchhoff’s laws, you need to decide which loops to use and the direction of current flow through each loop. In analyzing the circuit in [link], the direction of current flow was chosen to be clockwise, from point *a* to point *b*. How would the results change if the direction of the current was chosen to be counterclockwise, from point *b* to point *a*?

Solution:

The current calculated would be equal to $I = -0.20\text{ A}$ instead of $I = 0.20\text{ A}$. The sum of the power dissipated and the power consumed would still equal the power supplied.

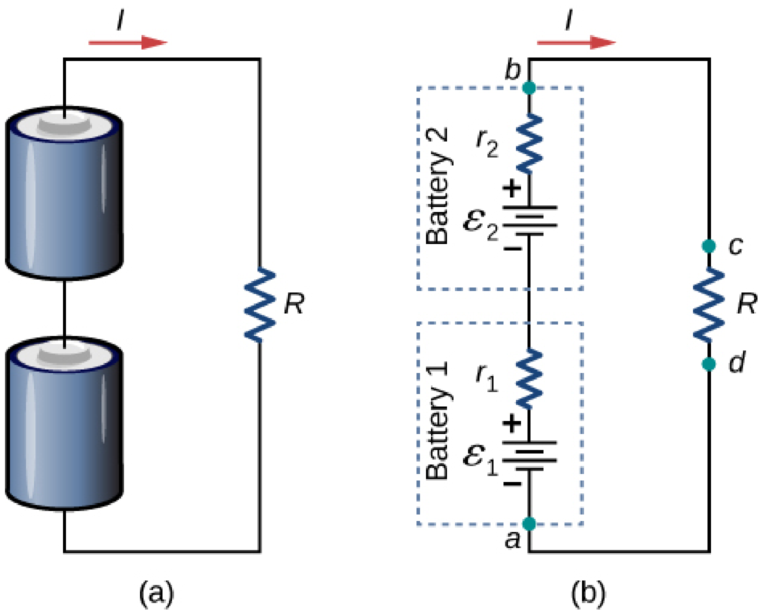
Multiple Voltage Sources

Many devices require more than one battery. Multiple voltage sources, such as batteries, can be connected in series configurations, parallel configurations, or a combination of the two.

In series, the positive terminal of one battery is connected to the negative terminal of another battery. Any number of voltage sources, including batteries, can be connected in series. Two batteries connected in series are shown in [link]. Using Kirchhoff’s loop rule for the circuit in part (b) gives the result

Equation:

$$\begin{aligned}&\varepsilon_1 - Ir_1 + \varepsilon_2 - Ir_2 - IR = 0,\\&\left[(\varepsilon_1 + \varepsilon_2) - I\left(r_1 + r_2\right)\right] - IR = 0.\end{aligned}$$



(a) Two batteries connected in series with a load resistor. (b) The circuit diagram of the two batteries and the load resistor, with each battery modeled as an idealized emf source and an internal resistance.

When voltage sources are in series, their internal resistances can be added together and their emfs can be added together to get the total values. Series connections of voltage sources are common—for example, in flashlights, toys, and other appliances. Usually, the cells are in series in order to produce a larger total emf. In [link], the terminal voltage is

Equation:

$$V_{\text{terminal}} = (\varepsilon_1 - Ir_1) + (\varepsilon_2 - Ir_2) = [(\varepsilon_1 + \varepsilon_2) - I(r_1 + r_2)] = (\varepsilon_1 + \varepsilon_2) + Ir_{\text{eq}}.$$

Note that the same current *I* is found in each battery because they are connected in series. The disadvantage of series connections of cells is that their internal resistances are additive.

Batteries are connected in series to increase the voltage supplied to the circuit. For instance, an LED flashlight may have two AAA cell batteries, each with a terminal voltage of 1.5 V, to provide 3.0 V to the flashlight.

Any number of batteries can be connected in series. For *N* batteries in series, the terminal voltage is equal to

Note:
Equation:

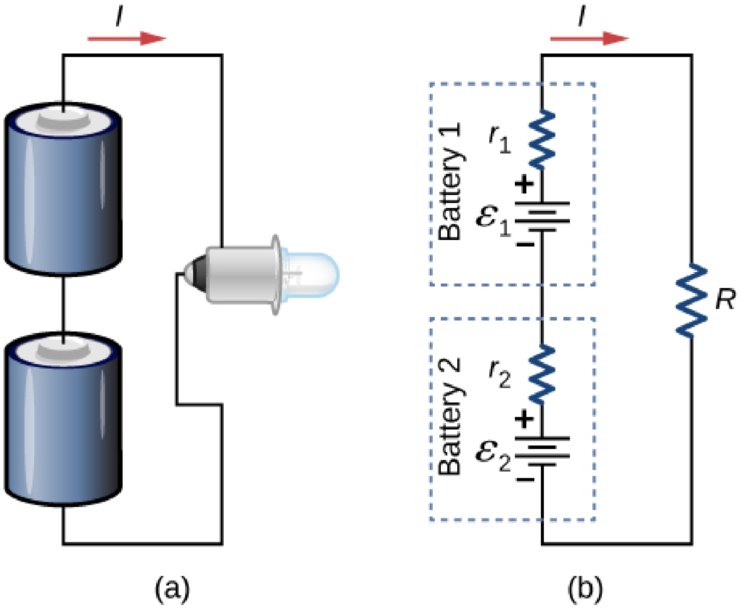
$$V_{\text{terminal}} = (\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{N-1} + \varepsilon_N) - I\,(r_1 + r_2 + \cdots + r_{N-1} + r_N) = \sum_{i=1}^{N} \varepsilon_i - Ir_{\text{eq}}$$

where the equivalent resistance is $r_{\text{eq}} = \sum_{i=1}^{N} r_i$.

When a load is placed across voltage sources in series, as in [link], we can find the current:
Equation:

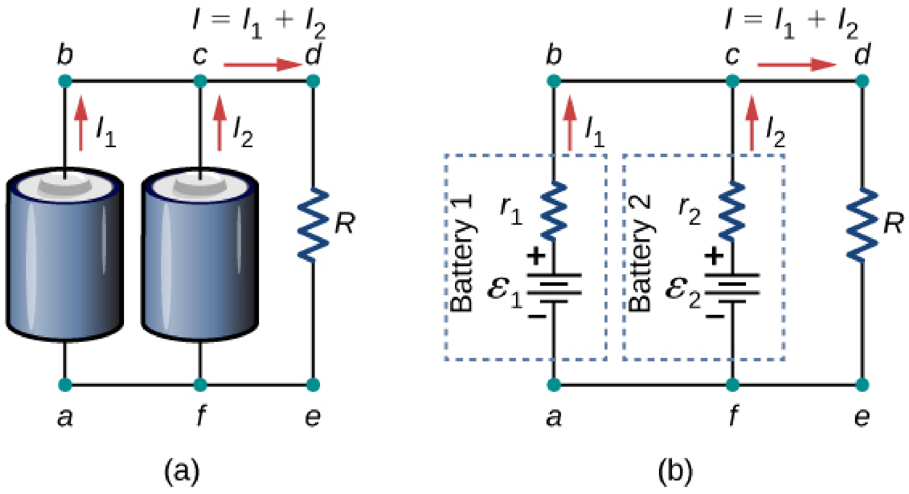
$$\begin{aligned}
&(\varepsilon_1 - Ir_1) + (\varepsilon_2 - Ir_2) = IR,\\
&Ir_1 + Ir_2 + IR = \varepsilon_1 + \varepsilon_2,\\
&I = \tfrac{\varepsilon_1 + \varepsilon_2}{r_1 + r_2 + R}.
\end{aligned}$$

As expected, the internal resistances increase the equivalent resistance.



Two batteries connect in series to an LED bulb, as found in a flashlight.

Voltage sources, such as batteries, can also be connected in parallel. [link] shows two batteries with identical emfs in parallel and connected to a load resistance. When the batteries are connect in parallel, the positive terminals are connected together and the negative terminals are connected together, and the load resistance is connected to the positive and negative terminals. Normally, voltage sources in parallel have identical emfs. In this simple case, since the voltage sources are in parallel, the total emf is the same as the individual emfs of each battery.



(a) Two batteries connect in parallel to a load resistor. (b) The circuit diagram shows the shows battery as an emf source and an internal resistor. The two emf sources have identical emfs (each labeled by ε) connected in parallel that produce the same emf.

Consider the Kirchhoff analysis of the circuit in [link](b). There are two loops and a node at point *b* and $\varepsilon = \varepsilon_1 = \varepsilon_2$.

Node *b*: $I_1 + I_2 - I = 0$.

Loop *abcfa*: $$\begin{aligned} \varepsilon - I_1 r_1 + I_2 r_2 - \varepsilon &= 0, \\ I_1 r_1 &= I_2 r_2. \end{aligned}$$

Loop *fcdef*: $$\begin{aligned} \varepsilon_2 - I_2 r_2 - IR &= 0, \\ \varepsilon - I_2 r_2 - IR &= 0. \end{aligned}$$

Solving for the current through the load resistor results in $I = \frac{\varepsilon}{r_{\text{eq}} + R}$, where $r_{\text{eq}} = \left(\frac{1}{r_1} + \frac{1}{r_2}\right)^{-1}$. The terminal voltage is equal to the potential drop across the load resistor $IR = \left(\frac{\varepsilon}{r_{\text{eq}} + R}\right)$. The parallel connection reduces the internal resistance and thus can produce a larger current.

Any number of batteries can be connected in parallel. For *N* batteries in parallel, the terminal voltage is equal to

Note:
Equation:

$$V_{\text{terminal}} = \varepsilon - I\left(\frac{1}{r_1} + \frac{1}{r_2} + \cdots + \frac{1}{r_{N-1}} + \frac{1}{r_N}\right)^{-1} = \varepsilon - Ir_{\text{eq}}$$

where the equivalent resistance is $r_{\text{eq}} = \left(\sum_{i=1}^{N} \frac{1}{r_i}\right)^{-1}$.

As an example, some diesel trucks use two 12-V batteries in parallel; they produce a total emf of 12 V but can deliver the larger current needed to start a diesel engine.

In summary, the terminal voltage of batteries in series is equal to the sum of the individual emfs minus the sum of the internal resistances times the current. When batteries are connected in parallel, they usually have equal emfs and the terminal voltage is equal to the emf minus the equivalent internal resistance times the current, where the equivalent internal resistance is smaller than the individual internal resistances. Batteries are connected in series to increase the terminal voltage to the load. Batteries are connected in parallel to increase the current to the load.

Solar Cell Arrays

Another example dealing with multiple voltage sources is that of combinations of solar cells—wired in both series and parallel combinations to yield a desired voltage and current. Photovoltaic generation, which is the conversion of sunlight directly into electricity, is based upon the photoelectric effect. The photoelectric effect is beyond the scope of this chapter and is covered in Photons and Matter Waves, but in general, photons hitting the surface of a solar cell create an electric current in the cell.

Most solar cells are made from pure silicon. Most single cells have a voltage output of about 0.5 V, while the current output is a function of the amount of sunlight falling on the cell (the incident solar radiation known as the insolation). Under bright noon sunlight, a current per unit area of about $100\ \text{mA/cm}^2$ of cell surface area is produced by typical single-crystal cells.

Individual solar cells are connected electrically in modules to meet electrical energy needs. They can be wired together in series or in parallel—connected like the batteries discussed earlier. A solar-cell array or module usually consists of between 36 and 72 cells, with a power output of 50 W to 140 W.

Solar cells, like batteries, provide a direct current (dc) voltage. Current from a dc voltage source is unidirectional. Most household appliances need an alternating current (ac) voltage.

Summary

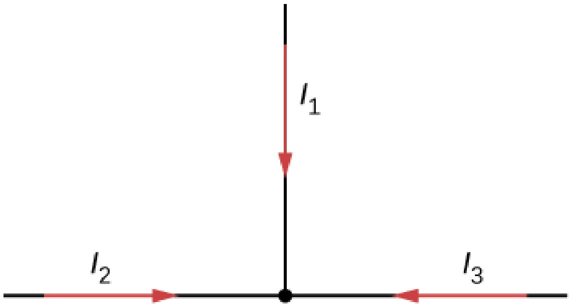
- Kirchhoff’s rules can be used to analyze any circuit, simple or complex. The simpler series and parallel connection rules are special cases of Kirchhoff’s rules.
- Kirchhoff’s first rule, also known as the junction rule, applies to the charge to a junction. Current is the flow of charge; thus, whatever charge flows into the junction must flow out.
- Kirchhoff’s second rule, also known as the loop rule, states that the voltage drop around a loop is zero.
- When calculating potential and current using Kirchhoff’s rules, a set of conventions must be followed for determining the correct signs of various terms.
- When multiple voltage sources are in series, their internal resistances add together and their emfs add together to get the total values.
- When multiple voltage sources are in parallel, their internal resistances combine to an equivalent resistance that is less than the individual resistance and provides a higher current than a single cell.
- Solar cells can be wired in series or parallel to provide increased voltage or current, respectively.

Conceptual Questions

Exercise:

Problem:

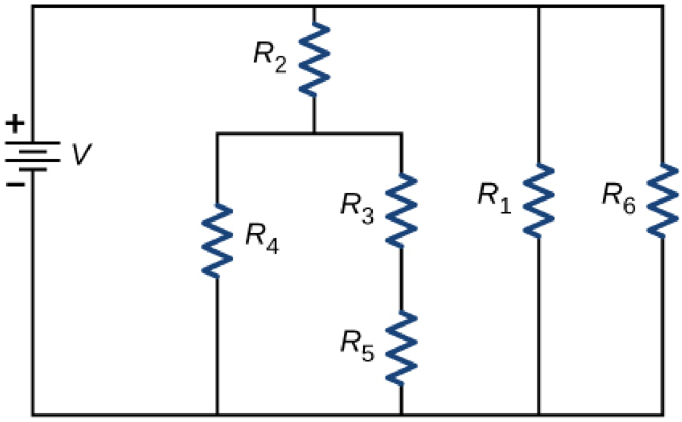
Can all of the currents going into the junction shown below be positive? Explain.



Exercise:

Problem:

Consider the circuit shown below. Does the analysis of the circuit require Kirchhoff’s method, or can it be redrawn to simplify the circuit? If it is a circuit of series and parallel connections, what is the equivalent resistance?



Solution:

It can be redrawn.

$$R_{\text{eq}} = \left[\frac{1}{R_6} + \frac{1}{R_1} + \frac{1}{R_2 + \left(\frac{1}{R_4} + \frac{1}{R_3 + R_5} \right)^{-1}} \right]^{-1}$$

Exercise:

Problem:

Do batteries in a circuit always supply power to a circuit, or can they absorb power in a circuit? Give an example.

Exercise:

Problem:

What are the advantages and disadvantages of connecting batteries in series? In parallel?

Solution:

In series the voltages add, but so do the internal resistances, because the internal resistances are in series. In parallel, the terminal voltage is the same, but the equivalent internal resistance is smaller than the smallest individual internal resistance and a higher current can be provided.

Exercise:

Problem:

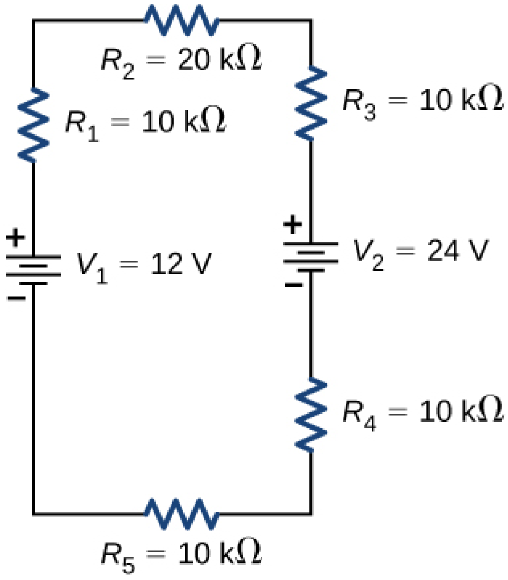
Semi-tractor trucks use four large 12-V batteries. The starter system requires 24 V, while normal operation of the truck's other electrical components utilizes 12 V. How could the four batteries be connected to produce 24 V? To produce 12 V? Why is 24 V better than 12 V for starting the truck's engine (a very heavy load)?

Problems

Exercise:

Problem:

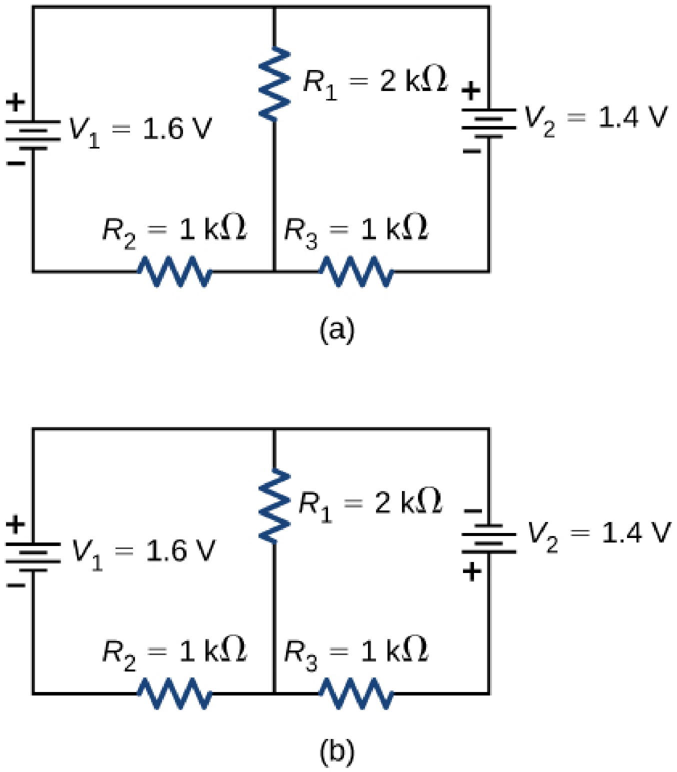
Consider the circuit shown below. (a) Find the voltage across each resistor. (b)What is the power supplied to the circuit and the power dissipated or consumed by the circuit?


$R_2 = 20\ \text{k}\Omega$
$R_1 = 10\ \text{k}\Omega$
$R_3 = 10\ \text{k}\Omega$
+
$V_1 = 12\ \text{V}$
−
+
$V_2 = 24\ \text{V}$
−
$R_4 = 10\ \text{k}\Omega$
$R_5 = 10\ \text{k}\Omega$

Exercise:

Problem:

Consider the circuits shown below. (a) What is the current through each resistor in part (a)? (b) What is the current through each resistor in part (b)? (c) What is the power dissipated or consumed by each circuit? (d) What is the power supplied to each circuit?

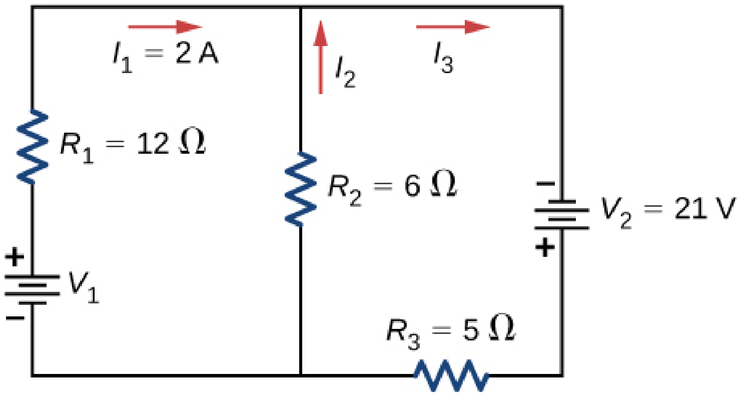

$R_1 = 2\ \text{k}\Omega$
$V_1 = 1.6\ \text{V}$
$V_2 = 1.4\ \text{V}$
$R_2 = 1\ \text{k}\Omega$
$R_3 = 1\ \text{k}\Omega$
(a)
$R_1 = 2\ \text{k}\Omega$
$V_1 = 1.6\ \text{V}$
$V_2 = 1.4\ \text{V}$
$R_2 = 1\ \text{k}\Omega$
$R_3 = 1\ \text{k}\Omega$
(b)

Solution:

a. $I_1 = 0.6\,\text{mA}, \quad I_2 = 0.4\,\text{mA}, \quad I_3 = 0.2\,\text{mA};$
b. $I_1 = 0.04\,\text{mA}, \quad I_2 = 1.52\,\text{mA}, \quad I_3 = -1.48\,\text{mA};$ c.
$P_{\text{out}} = 0.92\,\text{mW}, \quad P_{\text{out}} = 4.50\,\text{mW};$
d. $P_{\text{in}} = 0.92\,\text{mW}, \quad P_{\text{in}} = 4.50\,\text{mW}$

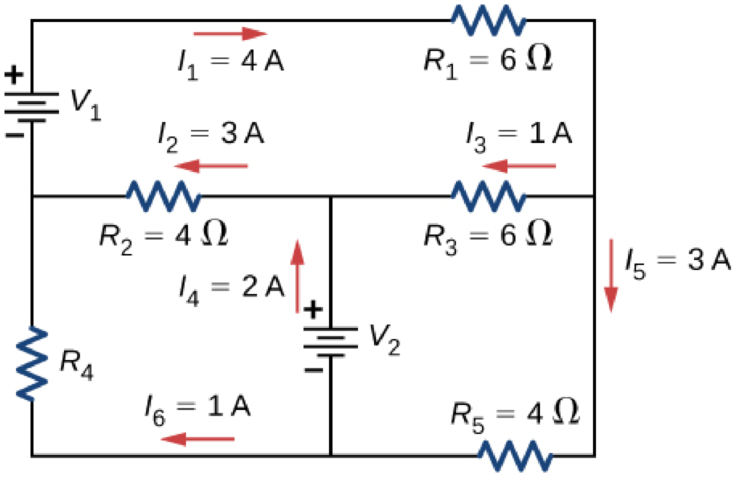
Exercise:

Problem: Consider the circuit shown below. Find $V_1, I_2,$ and I_3.


$I_1 = 2\ \text{A}$
I_2
I_3
$R_1 = 12\ \Omega$
$R_2 = 6\ \Omega$
$V_2 = 21\ \text{V}$
V_1
$R_3 = 5\ \Omega$

Exercise:

Problem: Consider the circuit shown below. Find $V_1, V_2,$ and R_4.

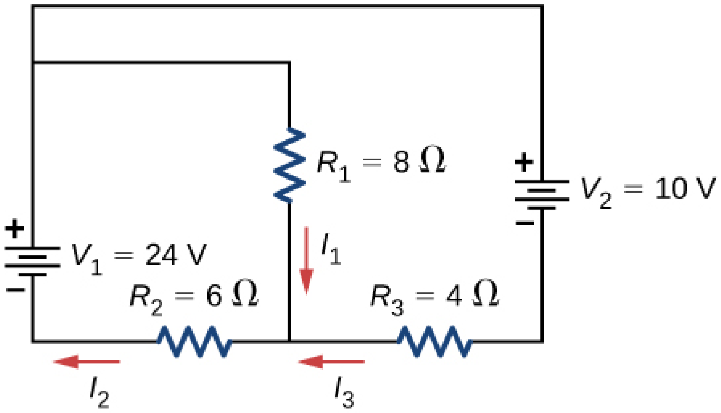

$I_1 = 4$ A
$R_1 = 6\ \Omega$
V_1
$I_2 = 3$ A
$I_3 = 1$ A
$R_2 = 4\ \Omega$
$R_3 = 6\ \Omega$
$I_5 = 3$ A
$I_4 = 2$ A
V_2
R_4
$I_6 = 1$ A
$R_5 = 4\ \Omega$

Solution:

$V_1 = 42\ \text{V}, V_2 = 6\ \text{V}, R_4 = 18\ \Omega$

Exercise:

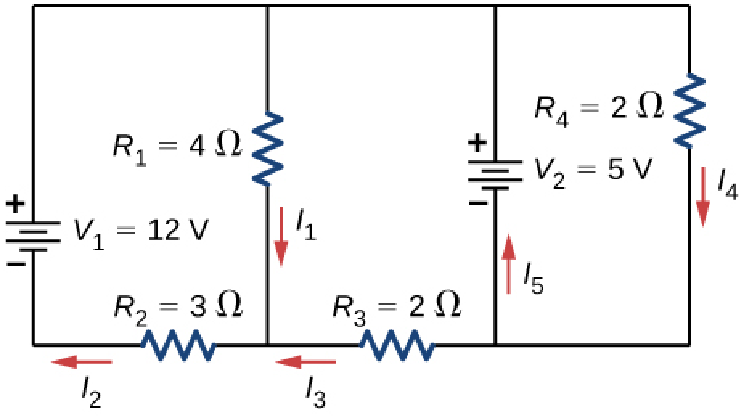
Problem: Consider the circuit shown below. Find $I_1, I_2,$ and I_3.


$R_1 = 8\ \Omega$
$V_2 = 10$ V
$V_1 = 24$ V
I_1
$R_2 = 6\ \Omega$
$R_3 = 4\ \Omega$
I_2
I_3

Exercise:

Problem:

Consider the circuit shown below. (a) Find $I_1, I_2, I_3, I_4,$ and I_5. (b) Find the power supplied by the voltage sources. (c) Find the power dissipated by the resistors.



$R_1 = 4\ \Omega$
$V_1 = 12$ V
$R_2 = 3\ \Omega$
$R_3 = 2\ \Omega$
$R_4 = 2\ \Omega$
$V_2 = 5$ V
I_1
I_2
I_3
I_4
I_5

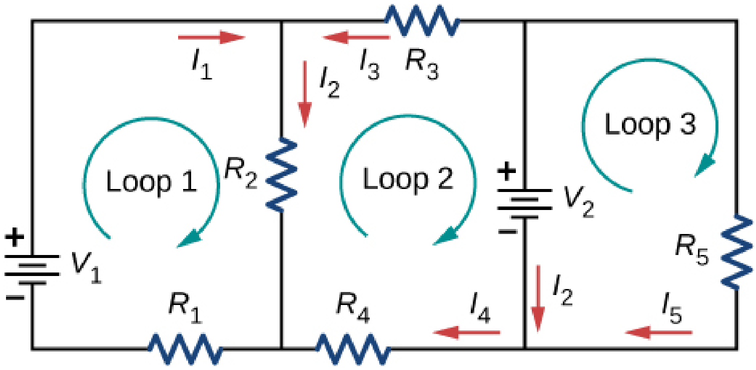
Solution:

a. $I_1 = 1.5\ \text{A}, I_2 = 2\ \text{A}, I_3 = 0.5\ \text{A},\ I_4 = 2.5\ \text{A}, I_5 = 2\ \text{A}$; b. $P_{\text{in}} = I_2V_1 + I_5V_5 = 34\ \text{W}$;
c. $P_{\text{out}} = I_1^2R_1 + I_2^2R_2 + I_3^2R_3 + I_4^2R_4 = 34\ \text{W}$

Exercise:

Problem:

Consider the circuit shown below. Write the three loop equations for the loops shown.

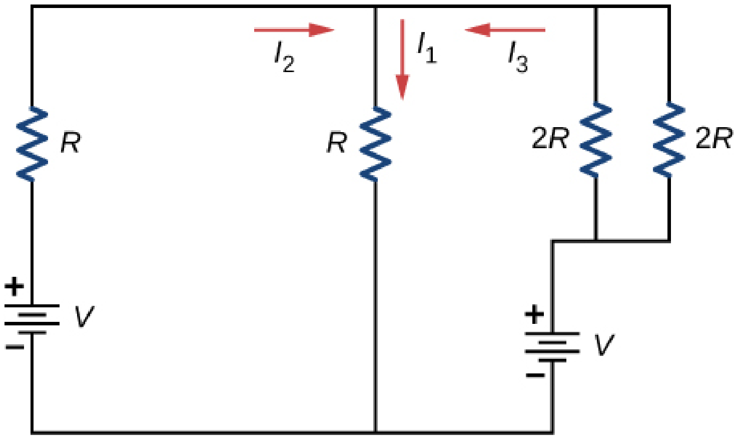


I_1
I_2
I_3
R_3
Loop 1
Loop 2
Loop 3
R_2
V_1
V_2
R_1
R_4
I_4
I_2
R_5
I_5

Exercise:

Problem:

Consider the circuit shown below. Write equations for the three currents in terms of R and V.



Solution:

$I_1 = \frac{2}{3}\frac{V}{R}, I_2 = \frac{V}{3R}, I_3 = \frac{V}{3R}$

Exercise:

Problem:

Consider the circuit shown in the preceding problem. Write equations for the power supplied by the voltage sources and the power dissipated by the resistors in terms of *R* and *V*.

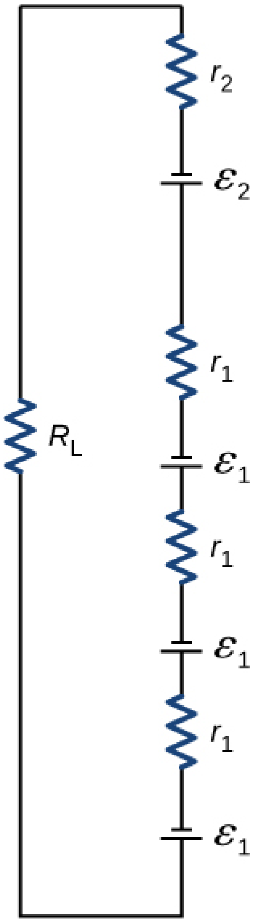
Exercise:

Problem:

A child’s electronic toy is supplied by three 1.58-V alkaline cells having internal resistances of $0.0200\ \Omega$ in series with a 1.53-V carbon-zinc dry cell having a $0.100\text{-}\Omega$ internal resistance. The load resistance is $10.0\ \Omega$. (a) Draw a circuit diagram of the toy and its batteries. (b) What current flows? (c) How much power is supplied to the load? (d) What is the internal resistance of the dry cell if it goes bad, resulting in only 0.500 W being supplied to the load?

Solution:

a.



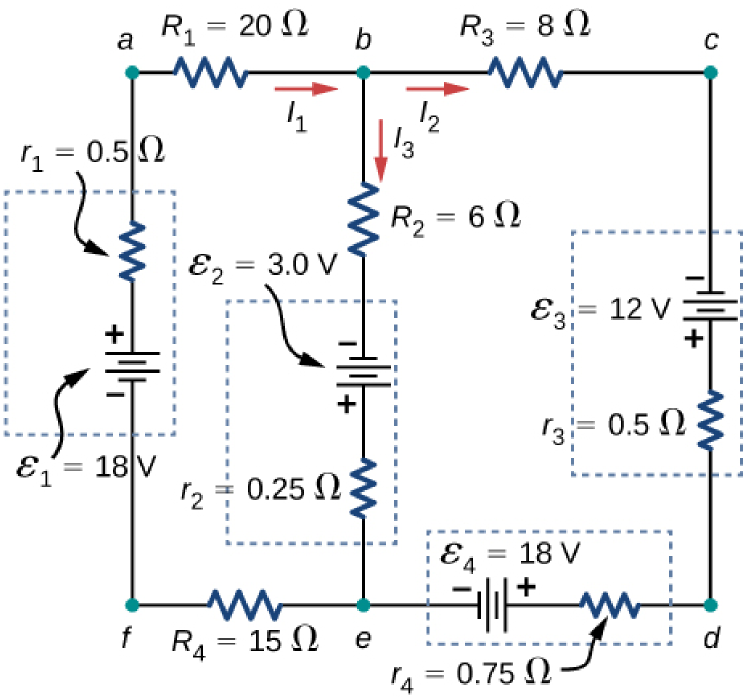
r_2
$\mathcal{E}_2$
r_1
R_L
$\mathcal{E}_1$
r_1
$\mathcal{E}_1$
r_1
$\mathcal{E}_1$

;
b. 0.617 A; c. 3.81 W; d. 18.0 Ω

Exercise:

Problem:

Apply the junction rule to Junction *b* shown below. Is any new information gained by applying the junction rule at *e*?


a
$R_1 = 20\ \Omega$
b
$R_3 = 8\ \Omega$
c
I_1
I_2
I_3
$r_1 = 0.5\ \Omega$
$R_2 = 6\ \Omega$
$\mathcal{E}_2 = 3.0$ V
$\mathcal{E}_3 = 12$ V
+
−
$r_3 = 0.5\ \Omega$
$\mathcal{E}_1 = 18$ V
$r_2 = 0.25\ \Omega$
$\mathcal{E}_4 = 18$ V
f
$R_4 = 15\ \Omega$
e
d
$r_4 = 0.75\ \Omega$

Exercise:

Problem: Apply the loop rule to Loop *afedcba* in the preceding problem.

Solution:

$$I_1r_1 - \varepsilon_1 + I_1R_4 + \varepsilon_4 + I_2r_4 + I_4r_3 - \varepsilon_3 + I_2R_3 + I_1R_1 = 0$$

Glossary

junction rule
: sum of all currents entering a junction must equal the sum of all currents leaving the junction

Kirchhoff’s rules
: set of two rules governing current and changes in potential in an electric circuit

loop rule
: algebraic sum of changes in potential around any closed circuit path (loop) must be zero

Electrical Measuring Instruments

By the end of the section, you will be able to:

- Describe how to connect a voltmeter in a circuit to measure voltage
- Describe how to connect an ammeter in a circuit to measure current
- Describe the use of an ohmmeter

Ohm’s law and Kirchhoff’s method are useful to analyze and design electrical circuits, providing you with the voltages across, the current through, and the resistance of the components that compose the circuit. To measure these parameters require instruments, and these instruments are described in this section.

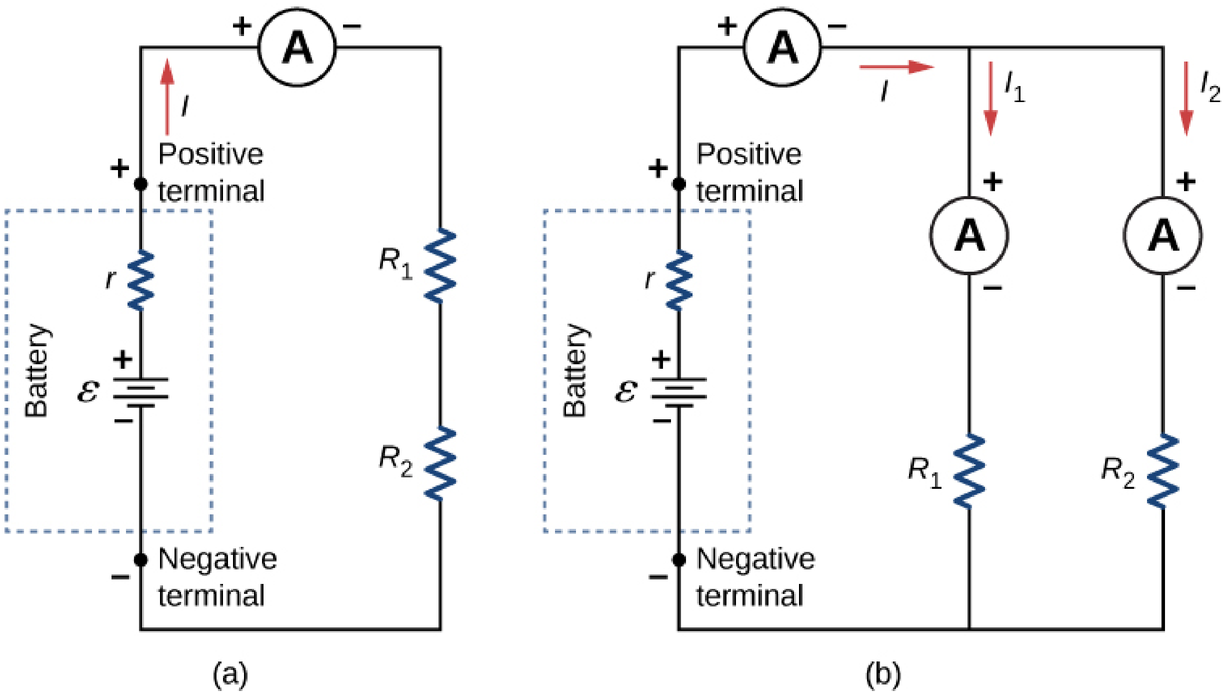
DC Voltmeters and Ammeters

Whereas **voltmeters** measure voltage, **ammeters** measure current. Some of the meters in automobile dashboards, digital cameras, cell phones, and tuner-amplifiers are actually voltmeters or ammeters ([link]). The internal construction of the simplest of these meters and how they are connected to the system they monitor give further insight into applications of series and parallel connections.

The fuel and temperature gauges (far right and far left, respectively) in this 1996 Volkswagen are voltmeters that register the voltage output of “sender” units. These units are proportional to the amount of gasoline in the tank and to the engine temperature. (credit: Christian Giersing)

Measuring Current with an Ammeter

To measure the current through a device or component, the ammeter is placed in series with the device or component. A series connection is used because objects in series have the same current passing through them. (See [link], where the ammeter is represented by the symbol A.)

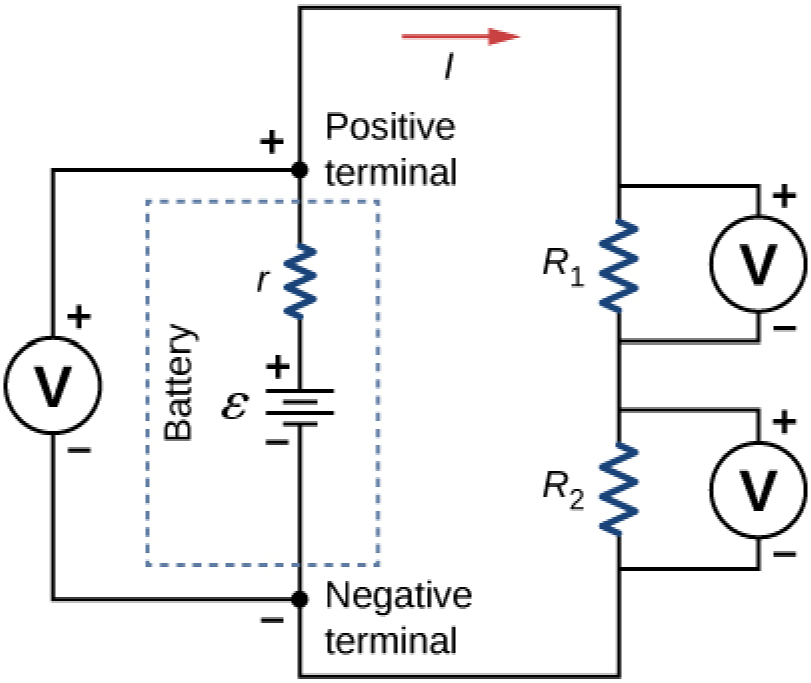


(a) When an ammeter is used to measure the current through two resistors connected in series to a battery, a single ammeter is placed in series with the two resistors because the current is the same through the two resistors in series. (b) When two resistors are connected in parallel with a battery, three meters, or three separate ammeter readings, are necessary to measure the current from the battery and through each resistor. The ammeter is connected in series with the component in question.

Ammeters need to have a very low resistance, a fraction of a milliohm. If the resistance is not negligible, placing the ammeter in the circuit would change the equivalent resistance of the circuit and modify the current that is being measured. Since the current in the circuit travels through the meter, ammeters normally contain a fuse to protect the meter from damage from currents which are too high.

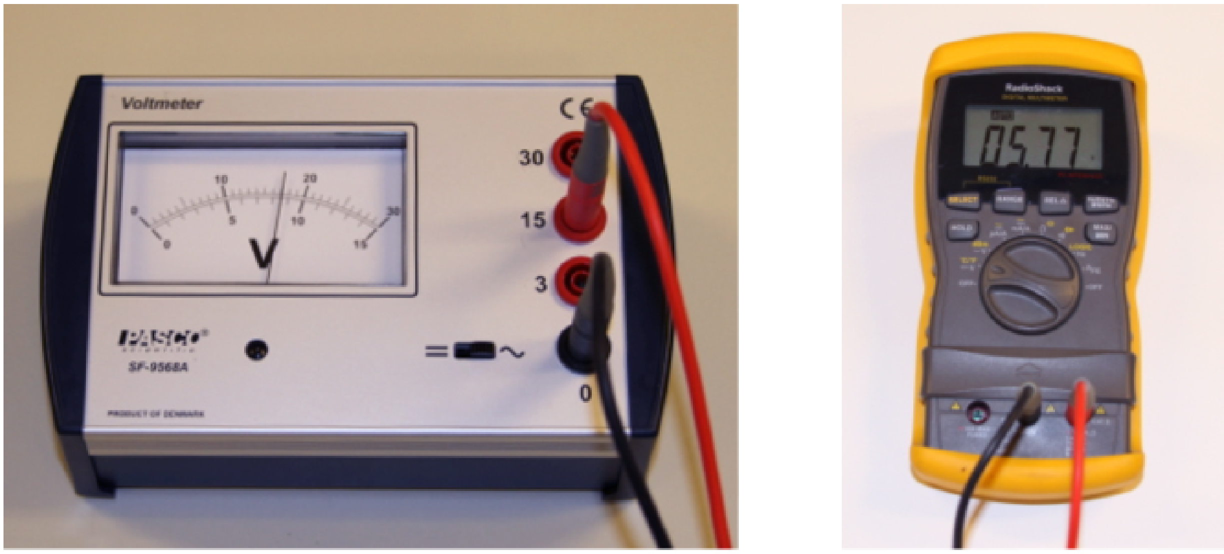
Measuring Voltage with a Voltmeter

A voltmeter is connected in parallel with whatever device it is measuring. A parallel connection is used because objects in parallel experience the same potential difference. (See [link], where the voltmeter is represented by the symbol V.)



To measure potential differences in this series circuit, the voltmeter (V) is placed in parallel with the voltage source or either of the resistors. Note that terminal voltage is measured between the positive terminal and the negative terminal of the battery or voltage source. It is not possible to connect a voltmeter directly across the emf without including the internal resistance *r* of the battery.

Since voltmeters are connected in parallel, the voltmeter must have a very large resistance. Digital voltmeters convert the analog voltage into a digital value to display on a digital readout ([link]). Inexpensive voltmeters have resistances on the order of $R_{\text{M}} = 10\,\text{M}\,\Omega$, whereas high-precision voltmeters have resistances on the order of $R_{\text{M}} = 10\,\text{G}\,\Omega$. The value of the resistance may vary, depending on which scale is used on the meter.



(a) (b)

(a) An analog voltmeter uses a galvanometer to measure the voltage. (b) Digital meters use an analog-to-digital converter to measure the voltage. (credit: modification of works by Joseph J. Trout)

Analog and Digital Meters

You may encounter two types of meters in the physics lab: analog and digital. The term ‘analog’ refers to signals or information represented by a continuously variable physical quantity, such as voltage or current. An analog meter uses a galvanometer, which is essentially a coil of wire with a small resistance, in a magnetic field, with a pointer attached that points to a scale. Current flows through the coil, causing the coil to rotate. To use the galvanometer as an ammeter, a small resistance is placed in parallel with the coil. For a voltmeter, a large resistance is placed in series with the coil. A digital meter uses a component called an analog-to-digital (A to D) converter and expresses the current or voltage as a series of the digits 0 and 1, which are used to run a digital display. Most analog meters have been replaced by digital meters.

Note:
Exercise:

Problem:

Check Your Understanding Digital meters are able to detect smaller currents than analog meters employing galvanometers. How does this explain their ability to measure voltage and current more accurately than analog meters?

Solution:

Since digital meters require less current than analog meters, they alter the circuit less than analog meters. Their resistance as a voltmeter can be far greater than an analog meter, and their resistance as an ammeter can be far less than an analog meter. Consult [link] and [link] and their discussion in the text.

Note:
In this virtual lab simulation, you may construct circuits with resistors, voltage sources, ammeters and voltmeters to test your knowledge of circuit design.

Ohmmeters

An ohmmeter is an instrument used to measure the resistance of a component or device. The operation of the ohmmeter is based on Ohm's law. Traditional ohmmeters contained an internal voltage source (such as a battery) that would be connected across the component to be tested, producing a current through the component. A galvanometer was then used to measure the current and the resistance was deduced using Ohm’s law. Modern digital meters use a constant current source to pass current through the component, and the voltage difference across the component is measured. In either case, the resistance is measured using Ohm’s law $(R = V/I)$, where the voltage is known and the current is measured, or the current is known and the voltage is measured.

The component of interest should be isolated from the circuit; otherwise, you will be measuring the equivalent resistance of the circuit. An ohmmeter should never be connected to a “live” circuit, one with a voltage source connected to it and current running through it. Doing so can damage the meter.

Summary

- Voltmeters measure voltage, and ammeters measure current. Analog meters are based on the combination of a resistor and a galvanometer, a device that gives an analog reading of current or voltage. Digital meters are based on analog-to-digital converters and provide a discrete or digital measurement of the current or voltage.
- A voltmeter is placed in parallel with the voltage source to receive full voltage and must have a large resistance to limit its effect on the circuit.
- An ammeter is placed in series to get the full current flowing through a branch and must have a small resistance to limit its effect on the circuit.
- Standard voltmeters and ammeters alter the circuit they are connected to and are thus limited in accuracy.
- Ohmmeters are used to measure resistance. The component in which the resistance is to be measured should be isolated (removed) from the circuit.

Conceptual Questions

Exercise:

Problem:

What would happen if you placed a voltmeter in series with a component to be tested?

Solution:

The voltmeter would put a large resistance in series with the circuit, significantly changing the circuit. It would probably give a reading, but it would be meaningless.

Exercise:

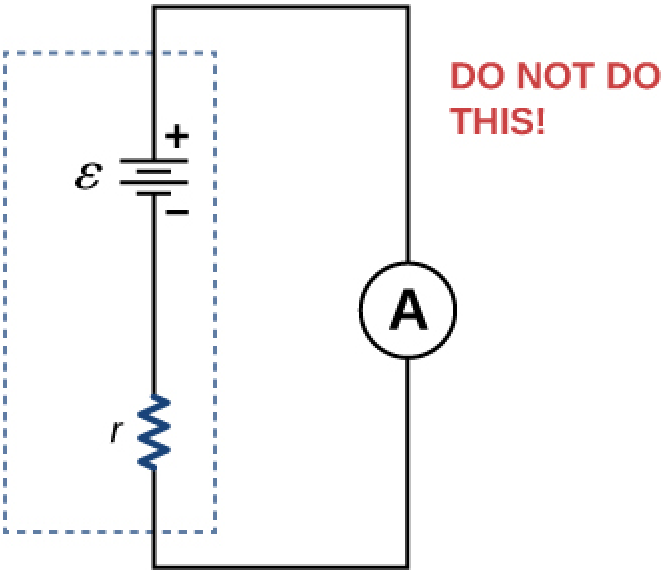
Problem:

What is the basic operation of an ohmmeter as it measures a resistor?

Exercise:

Problem:

Why should you not connect an ammeter directly across a voltage source as shown below?



Solution:

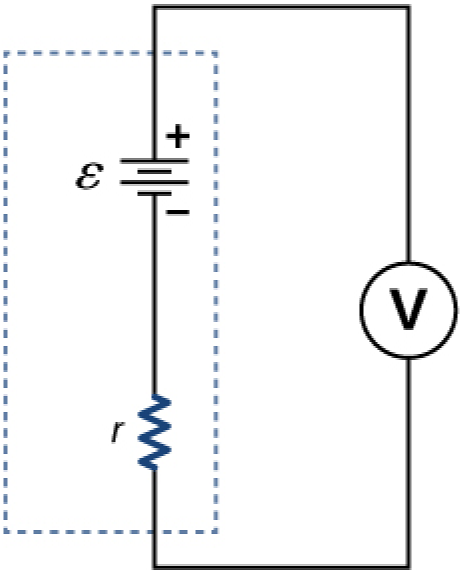
The ammeter has a small resistance; therefore, a large current will be produced and could damage the meter and/or overheat the battery.

Problems

Exercise:

Problem:

Suppose you measure the terminal voltage of a 1.585-V alkaline cell having an internal resistance of 0.100 Ω by placing a 1.00-kΩ voltmeter across its terminals (see below). (a) What current flows? (b) Find the terminal voltage. (c) To see how close the measured terminal voltage is to the emf, calculate their ratio.



Glossary

ammeter
    instrument that measures current

voltmeter
    instrument that measures voltage

RC Circuits
By the end of the section, you will be able to:

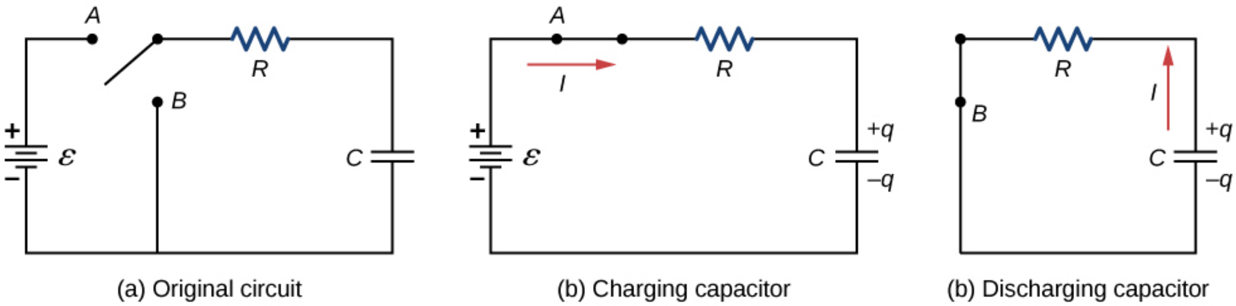
- Describe the charging process of a capacitor
- Describe the discharging process of a capacitor
- List some applications of RC circuits

When you use a flash camera, it takes a few seconds to charge the capacitor that powers the flash. The light flash discharges the capacitor in a tiny fraction of a second. Why does charging take longer than discharging? This question and several other phenomena that involve charging and discharging capacitors are discussed in this module.

Circuits with Resistance and Capacitance

An ***RC* circuit** is a circuit containing resistance and capacitance. As presented in Capacitance, the capacitor is an electrical component that stores electric charge, storing energy in an electric field.

[link](a) shows a simple *RC* circuit that employs a dc (direct current) voltage source ε, a resistor *R*, a capacitor *C*, and a two-position switch. The circuit allows the capacitor to be charged or discharged, depending on the position of the switch. When the switch is moved to position *A*, the capacitor charges, resulting in the circuit in part (b). When the switch is moved to position *B*, the capacitor discharges through the resistor.



(a) Original circuit (b) Charging capacitor (b) Discharging capacitor

(a) An *RC* circuit with a two-pole switch that can be used to charge and discharge a capacitor. (b) When the switch is moved to position *A*, the circuit reduces to a simple series connection of the voltage source, the resistor, the capacitor, and the switch. (c) When the switch is moved to position *B*, the circuit reduces to a simple series connection of the resistor, the capacitor, and the switch. The voltage source is removed from the circuit.

Charging a Capacitor

We can use Kirchhoff's loop rule to understand the charging of the capacitor. This results in the equation $\varepsilon - V_R - V_c = 0$. This equation can be used to model the charge as a function of time as the capacitor charges. Capacitance is defined as $C = q/V$, so the voltage across the capacitor is $V_C = \frac{q}{C}$. Using Ohm's law, the potential drop across the resistor is $V_R = IR$, and the current is defined as $I = dq/dt$.
Equation:

$$\begin{array}{l}\varepsilon - V_R - V_c = 0, \\ \varepsilon - IR - \frac{q}{C} = 0, \\ \varepsilon - R\frac{dq}{dt} - \frac{q}{C} = 0.\end{array}$$

This differential equation can be integrated to find an equation for the charge on the capacitor as a function of time.

Equation:

$$\begin{array}{l}\varepsilon - R\frac{dq}{dt} - \frac{q}{C} = 0, \\ \frac{dq}{dt} = \frac{\varepsilon C - q}{RC}, \\ \int\limits_0^q \frac{dq}{\varepsilon C - q} = \frac{1}{RC}\int\limits_0^t dt.\end{array}$$

Let $u = \varepsilon C - q$, then $du = -dq$. The result is

Equation:

$$\begin{array}{l}-\int\limits_0^q \frac{du}{u} = \frac{1}{RC}\int\limits_0^t dt, \\ \ln\left(\frac{\varepsilon C - q}{\varepsilon C}\right) = -\frac{1}{RC}t, \\ \frac{\varepsilon C - q}{\varepsilon C} = e^{-t/RC}.\end{array}$$

Simplifying results in an equation for the charge on the charging capacitor as a function of time:

Note:

Equation:

$$q\left(t\right) = C\varepsilon\left(1 - e^{-\frac{t}{RC}}\right) = Q\left(1 - e^{-\frac{t}{\tau}}\right).$$

A graph of the charge on the capacitor versus time is shown in [link](a). First note that as time approaches infinity, the exponential goes to zero, so the charge approaches the maximum charge $Q = C\varepsilon$ and has units of coulombs. The units of *RC* are seconds, units of time. This quantity is known as the time constant:

Note:

Equation:

$$\tau = RC.$$

At time $t = \tau = RC$, the charge is equal to $1 - e^{-1} = 1 - 0.368 = 0.632$ of the maximum charge $Q = C\varepsilon$. Notice that the time rate change of the charge is the slope at a point of the charge versus time plot. The slope of the graph is large at time $t = 0.0$ s and approaches zero as time increases.

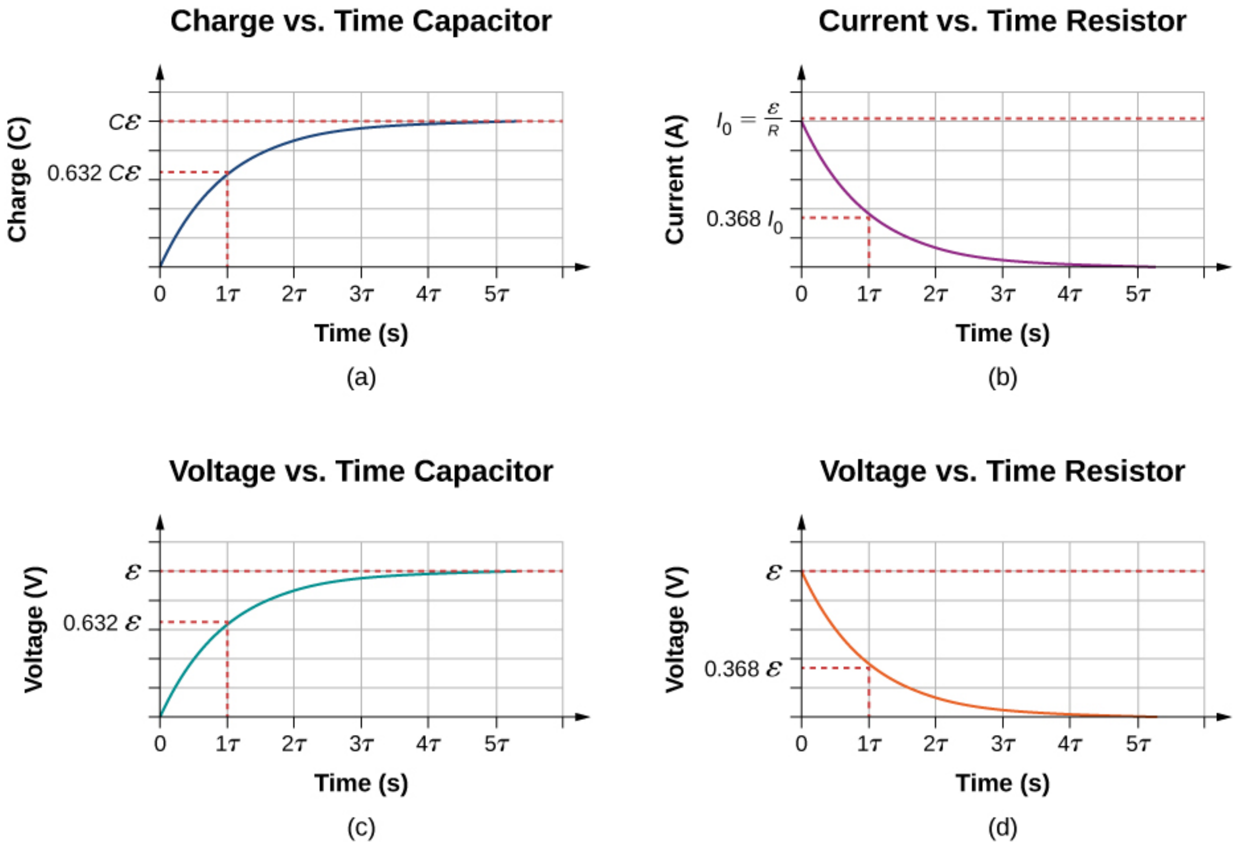
As the charge on the capacitor increases, the current through the resistor decreases, as shown in [link](b). The current through the resistor can be found by taking the time derivative of the charge.
Equation:

$$I\left(t\right) = \frac{dq}{dt} = \frac{d}{dt}\left[C\varepsilon\left(1 - e^{-\frac{t}{RC}}\right)\right],$$
$$I\left(t\right) = C\varepsilon\left(\frac{1}{RC}\right)e^{-\frac{t}{RC}} = \frac{\varepsilon}{R}e^{-\frac{t}{RC}} = I_o e^{-\frac{t}{RC}},$$

Note:
Equation:

$$I\left(t\right) = I_0 e^{-t/\tau}.$$

At time $t = 0.00$ s, the current through the resistor is $I_0 = \frac{\varepsilon}{R}$. As time approaches infinity, the current approaches zero. At time $t = \tau$, the current through the resistor is $I\left(t = \tau\right) = I_0 e^{-1} = 0.368 I_0$.



(a) Charge on the capacitor versus time as the capacitor charges. (b) Current through the resistor versus time. (c) Voltage difference across the capacitor. (d) Voltage difference across the resistor.

[link](c) and [link](d) show the voltage differences across the capacitor and the resistor, respectively. As the charge on the capacitor increases, the current decreases, as does the voltage difference across the resistor $V_R\left(t\right) = (I_0R)e^{-t/\tau} = \varepsilon e^{-t/\tau}$. The voltage difference across the capacitor increases as $V_C\left(t\right) = \varepsilon\left(1 - e^{-t/\tau}\right)$.

Discharging a Capacitor

When the switch in [link](a) is moved to position B, the circuit reduces to the circuit in part (c), and the charged capacitor is allowed to discharge through the resistor. A graph of the charge on the capacitor as a function of time is shown in [link](a). Using Kirchhoff's loop rule to analyze the circuit as the capacitor discharges results in the equation $-V_R - V_c = 0$, which simplifies to $IR + \frac{q}{C} = 0$. Using the definition of current $\frac{dq}{dt}R = -\frac{q}{C}$ and integrating the loop equation yields an equation for the charge on the capacitor as a function of time:

Note:
Equation:

$$q\left(t\right) = Qe^{-t/\tau}.$$

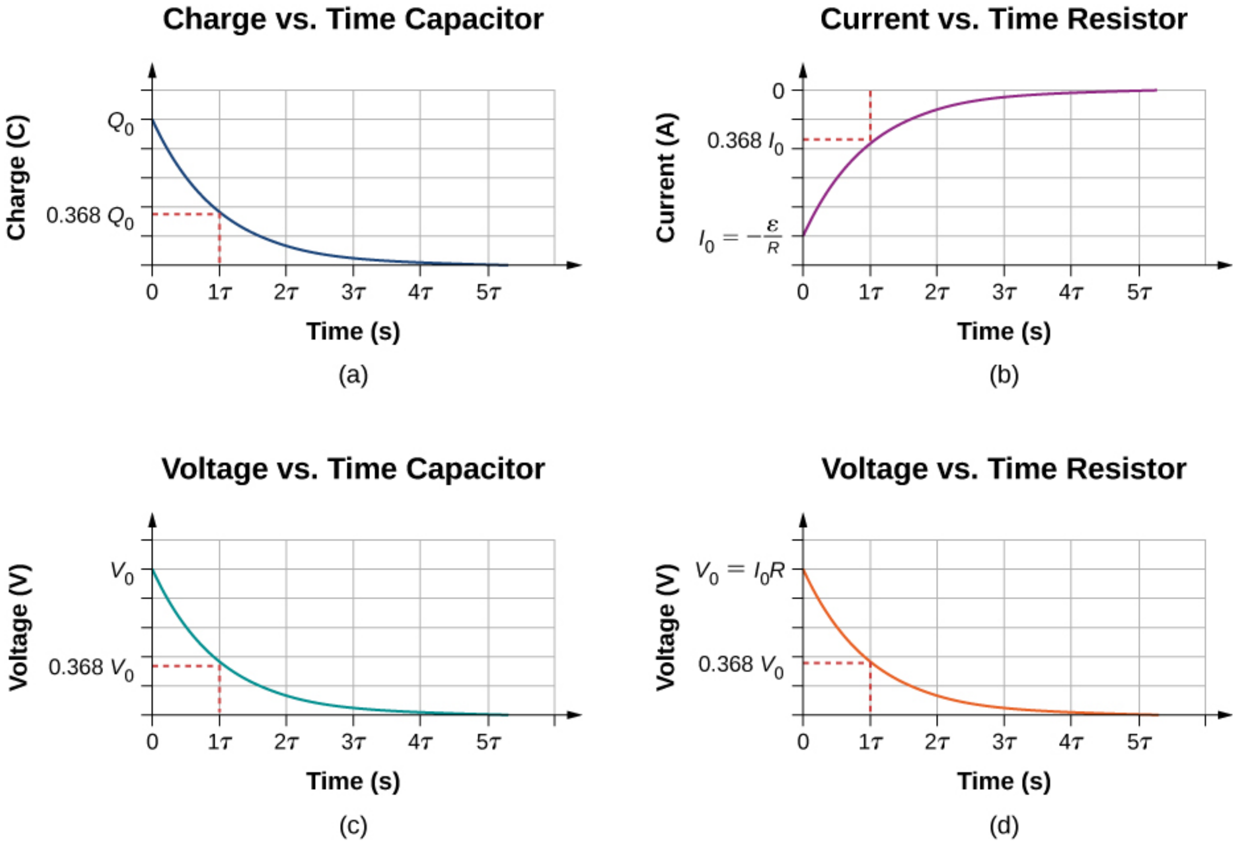
Here, Q is the initial charge on the capacitor and $\tau = RC$ is the time constant of the circuit. As shown in the graph, the charge decreases exponentially from the initial charge, approaching zero as time approaches infinity.

The current as a function of time can be found by taking the time derivative of the charge:

Note:
Equation:

$$I\left(t\right) = -\frac{Q}{RC}e^{-t/\tau}.$$

The negative sign shows that the current flows in the opposite direction of the current found when the capacitor is charging. [link](b) shows an example of a plot of charge versus time and current versus time. A plot of the voltage difference across the capacitor and the voltage difference across the resistor as a function of time are shown in parts (c) and (d) of the figure. Note that the magnitudes of the charge, current, and voltage all decrease exponentially, approaching zero as time increases.



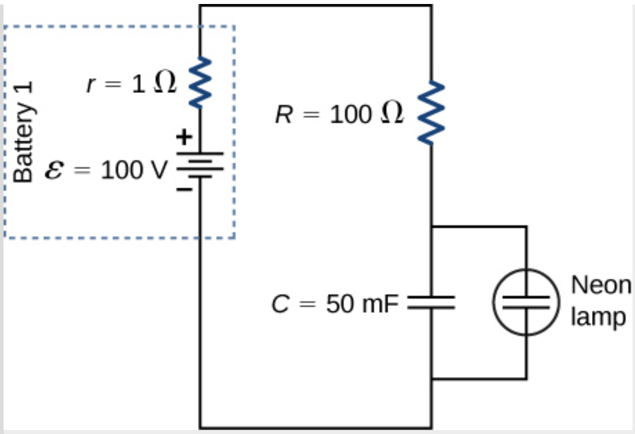
(a) Charge on the capacitor versus time as the capacitor discharges. (b) Current through the resistor versus time. (c) Voltage difference across the capacitor. (d) Voltage difference across the resistor.

Now we can explain why the flash camera mentioned at the beginning of this section takes so much longer to charge than discharge: The resistance while charging is significantly greater than while discharging. The internal resistance of the battery accounts for most of the resistance while charging. As the battery ages, the increasing internal resistance makes the charging process even slower.

Example:

The Relaxation Oscillator

One application of an *RC* circuit is the relaxation oscillator, as shown below. The relaxation oscillator consists of a voltage source, a resistor, a capacitor, and a neon lamp. The neon lamp acts like an open circuit (infinite resistance) until the potential difference across the neon lamp reaches a specific voltage. At that voltage, the lamp acts like a short circuit (zero resistance), and the capacitor discharges through the neon lamp and produces light. In the relaxation oscillator shown, the voltage source charges the capacitor until the voltage across the capacitor is 80 V. When this happens, the neon in the lamp breaks down and allows the capacitor to discharge through the lamp, producing a bright flash. After the capacitor fully discharges through the neon lamp, it begins to charge again, and the process repeats. Assuming that the time it takes the capacitor to discharge is negligible, what is the time interval between flashes?



Strategy

The time period can be found from considering the equation $V_C\left(t\right) = \varepsilon\left(1 - e^{-t/\tau}\right)$, where$\tau = (R + r)C$.

Solution

The neon lamp flashes when the voltage across the capacitor reaches 80 V. The *RC* time constant is equal to $\tau = (R + r)C = (101\ \Omega)\left(50\ \times\ 10^{-3}\text{F}\right) = 5.05\ \text{s}$. We can solve the voltage equation for the time it takes the capacitor to reach 80 V:

Equation:

$$\begin{aligned} V_C\left(t\right) &= \varepsilon\left(1 - e^{-t/\tau}\right), \\ e^{-t/\tau} &= 1 - \frac{V_C(t)}{\varepsilon}, \\ \ln\left(e^{-t/\tau}\right) &= \ln\left(1 - \frac{V_C(t)}{\varepsilon}\right), \\ t &= -\tau\ln\left(1 - \frac{V_C(t)}{\varepsilon}\right) = -5.05\ \text{s} \cdot \ln\left(1 - \frac{80\ \text{V}}{100\ \text{V}}\right) = 8.13\ \text{s}. \end{aligned}$$

Significance

One application of the relaxation oscillator is for controlling indicator lights that flash at a frequency determined by the values for *R* and *C*. In this example, the neon lamp will flash every 8.13 seconds, a frequency of $f = \frac{1}{T} = \frac{1}{8.13\ \text{s}} = 0.123\ \text{Hz}$. The relaxation oscillator has many other practical uses. It is often used in electronic circuits, where the neon lamp is replaced by a transistor or a device known as a tunnel diode. The description of the transistor and tunnel diode is beyond the scope of this chapter, but you can think of them as voltage controlled switches. They are normally open switches, but when the right voltage is applied, the switch closes and conducts. The “switch” can be used to turn on another circuit, turn on a light, or run a small motor. A relaxation oscillator can be used to make the turn signals of your car blink or your cell phone to vibrate.

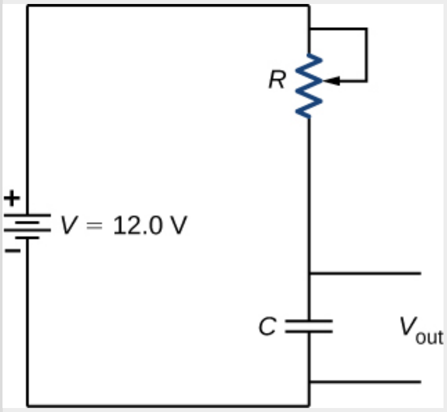
RC circuits have many applications. They can be used effectively as timers for applications such as intermittent windshield wipers, pace makers, and strobe lights. Some models of intermittent windshield wipers use a variable resistor to adjust the interval between sweeps of the wiper. Increasing the resistance increases the *RC* time constant, which increases the time between the operation of the wipers.

Another application is the pacemaker. The heart rate is normally controlled by electrical signals, which cause the muscles of the heart to contract and pump blood. When the heart rhythm is abnormal (the heartbeat is too high or too low), pace makers can be used to correct this abnormality. Pacemakers have sensors that detect body motion and breathing to increase the heart rate during physical activities, thus meeting the increased need for blood and oxygen, and an *RC* timing circuit can be used to control the time between voltage signals to the heart.

Looking ahead to the study of ac circuits (Alternating-Current Circuits), ac voltages vary as sine functions with specific frequencies. Periodic variations in voltage, or electric signals, are often recorded by scientists. These voltage signals could come from music recorded by a microphone or atmospheric data collected by radar. Occasionally, these signals can contain unwanted frequencies known as “noise.” *RC* filters can be used to filter out the unwanted frequencies.

In the study of electronics, a popular device known as a 555 timer provides timed voltage pulses. The time between pulses is controlled by an *RC* circuit. These are just a few of the countless applications of *RC* circuits.

Example:
Intermittent Windshield Wipers
A relaxation oscillator is used to control a pair of windshield wipers. The relaxation oscillator consists of a 10.00-mF capacitor and a 10.00-kΩ variable resistor known as a rheostat. A knob connected to the variable resistor allows the resistance to be adjusted from $0.00\ \Omega$ to $10.00\ \text{k}\Omega$. The output of the capacitor is used to control a voltage-controlled switch. The switch is normally open, but when the output voltage reaches 10.00 V, the switch closes, energizing an electric motor and discharging the capacitor. The motor causes the windshield wipers to sweep once across the windshield and the capacitor begins to charge again. To what resistance should the rheostat be adjusted for the period of the wiper blades be 10.00 seconds?



Strategy
The resistance considers the equation $V_{\text{out}}\left(t\right)=V\left(1-e^{-t/\tau}\right)$, where $\tau=RC$. The capacitance, output voltage, and voltage of the battery are given. We need to solve this equation for the resistance.
Solution
The output voltage will be 10.00 V and the voltage of the battery is 12.00 V. The capacitance is given as 10.00 mF. Solving for the resistance yields
Equation:

$$\begin{aligned}
V_{\text{out}}\left(t\right) &= V\left(1-e^{-t/\tau}\right),\\
e^{-t/RC} &= 1-\frac{V_{\text{out}}(t)}{V},\\
\ln\left(e^{-t/RC}\right) &= \ln\left(1-\frac{V_{\text{out}}(t)}{V}\right),\\
-\frac{t}{RC} &= \ln\left(1-\frac{V_{\text{out}}(t)}{V}\right),\\
R &= \frac{-t}{C\ln\left(1-\frac{V_C(t)}{V}\right)}=\frac{-10.00\ \text{s}}{10\times10^{-3}\ \text{F}\ln\left(1-\frac{10\ \text{V}}{12\ \text{V}}\right)}=558.11\ \Omega.
\end{aligned}$$

Significance

Increasing the resistance increases the time delay between operations of the windshield wipers. When the resistance is zero, the windshield wipers run continuously. At the maximum resistance, the period of the operation of the wipers is:

Equation:

$$t=-RC\ln\left(1-\frac{V_{\text{out}}\left(t\right)}{V}\right)=-\left(10\ \times\ 10^{-3}\ \text{F}\right)\left(10\ \times\ 10^{3}\ \Omega\right)\ln\left(1-\frac{10\ \text{V}}{12\ \text{V}}\right)=179.18\ \text{s}=2.98\ \text{min}.$$

The *RC* circuit has thousands of uses and is a very important circuit to study. Not only can it be used to time circuits, it can also be used to filter out unwanted frequencies in a circuit and used in power supplies, like the one for your computer, to help turn ac voltage to dc voltage.

Summary

- An *RC* circuit is one that has both a resistor and a capacitor.
- The time constant τ for an *RC* circuit is $\tau = RC$.
- When an initially uncharged $(q=0\ \text{at}\ t=0)$ capacitor in series with a resistor is charged by a dc voltage source, the capacitor asymptotically approaches the maximum charge.
- As the charge on the capacitor increases, the current exponentially decreases from the initial current: $I_0=\varepsilon/R$.
- If a capacitor with an initial charge Q is discharged through a resistor starting at $t=0$, then its charge decreases exponentially. The current flows in the opposite direction, compared to when it charges, and the magnitude of the charge decreases with time.

Conceptual Questions

Exercise:

Problem:

A battery, switch, capacitor, and lamp are connected in series. Describe what happens to the lamp when the switch is closed.

Exercise:

Problem:

When making an ECG measurement, it is important to measure voltage variations over small time intervals. The time is limited by the *RC* constant of the circuit—it is not possible to measure time variations shorter than *RC*. How would you manipulate *R* and *C* in the circuit to allow the necessary measurements?

Solution:

The time constant can be shortened by using a smaller resistor and/or a smaller capacitor. Care should be taken when reducing the resistance because the initial current will increase as the resistance decreases.

Problems

Exercise:

Problem:

The timing device in an automobile’s intermittent wiper system is based on an *RC* time constant and utilizes a $0.500\text{-}\mu\text{F}$ capacitor and a variable resistor. Over what range must *R* be made to vary to achieve time constants from 2.00 to 15.0 s?

Solution:

$4.00\text{ to }30.0\text{ M}\Omega$

Exercise:

Problem:

A heart pacemaker fires 72 times a minute, each time a 25.0-nF capacitor is charged (by a battery in series with a resistor) to 0.632 of its full voltage. What is the value of the resistance?

Exercise:

Problem:

The duration of a photographic flash is related to an *RC* time constant, which is $0.100 \mu\text{s}$ for a certain camera. (a) If the resistance of the flash lamp is $0.0400\ \Omega$ during discharge, what is the size of the capacitor supplying its energy? (b) What is the time constant for charging the capacitor, if the charging resistance is $800\text{ k}\Omega$?

Solution:

a. $2.50\ \mu\text{F}$; b. 2.00 s

Exercise:

Problem:

A 2.00- and a $7.50\text{-}\mu\text{F}$ capacitor can be connected in series or parallel, as can a 25.0- and a $100\text{-k}\Omega$ resistor. Calculate the four *RC* time constants possible from connecting the resulting capacitance and resistance in series.

Exercise:

Problem:

A $500\text{-}\Omega$ resistor, an uncharged $1.50\text{-}\mu\text{F}$ capacitor, and a 6.16-V emf are connected in series. (a) What is the initial current? (b) What is the *RC* time constant? (c) What is the current after one time constant? (d) What is the voltage on the capacitor after one time constant?

Solution:

a. 12.3 mA; b. $7.50\ \times\ 10^{-4}\text{s}$; c. 4.53 mA; d. 3.89 V

Exercise:

Problem:

A heart defibrillator being used on a patient has an *RC* time constant of 10.0 ms due to the resistance of the patient and the capacitance of the defibrillator. (a) If the defibrillator has a capacitance of $8.00 \mu\text{F}$, what is the resistance of the path through the patient? (You may neglect the capacitance of the patient and the resistance of the defibrillator.) (b) If the initial voltage is 12.0 kV, how long does it take to decline to $6.00\ \times\ 10^{2}\text{ V}$?

Exercise:

Problem:

An ECG monitor must have an *RC* time constant less than $1.00 \times 10^2 \mu\text{s}$ to be able to measure variations in voltage over small time intervals. (a) If the resistance of the circuit (due mostly to that of the patient's chest) is $1.00\,\text{k}\Omega$, what is the maximum capacitance of the circuit? (b) Would it be difficult in practice to limit the capacitance to less than the value found in (a)?

Solution:

a. $1.00 \times 10^{-7}\,\text{F}$; b. No, in practice it would not be difficult to limit the capacitance to less than 100 nF, since typical capacitors range from fractions of a picofarad (pF) to milifarad (mF).

Exercise:

Problem:

Using the exact exponential treatment, determine how much time is required to charge an initially uncharged 100-pF capacitor through a $75.0\text{-M}\Omega$ resistor to 90.0 of its final voltage.

Exercise:

Problem:

If you wish to take a picture of a bullet traveling at 500 m/s, then a very brief flash of light produced by an *RC* discharge through a flash tube can limit blurring. Assuming 1.00 mm of motion during one *RC* constant is acceptable, and given that the flash is driven by a $600\text{-}\mu\text{F}$ capacitor, what is the resistance in the flash tube?

Solution:

$3.33 \times 10^{-3}\,\Omega$

Glossary

RC circuit
: circuit that contains both a resistor and a capacitor

Household Wiring and Electrical Safety
By the end of the section, you will be able to:

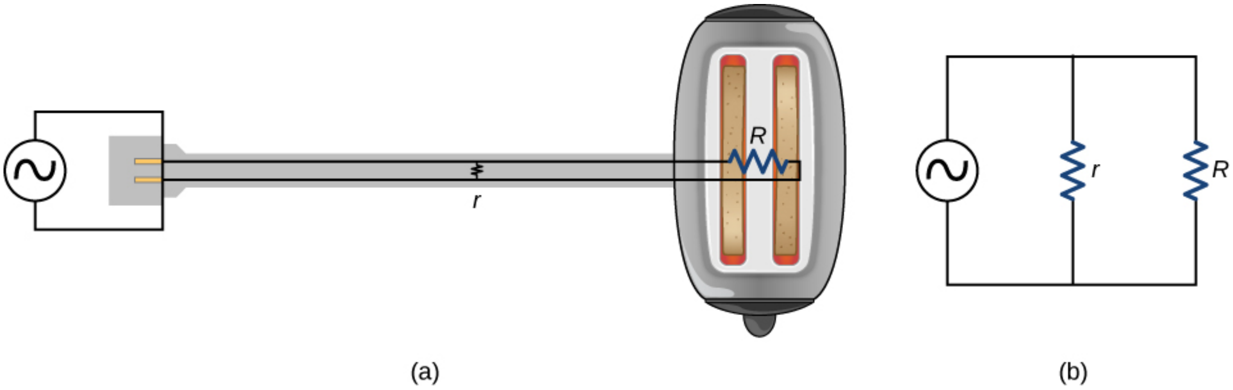
- List the basic concepts involved in house wiring
- Define the terms thermal hazard and shock hazard
- Describe the effects of electrical shock on human physiology and their relationship to the amount of current through the body
- Explain the function of fuses and circuit breakers

Electricity presents two known hazards: thermal and shock. A **thermal hazard** is one in which an excessive electric current causes undesired thermal effects, such as starting a fire in the wall of a house. A **shock hazard** occurs when an electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. In this section, we consider these hazards and the various factors affecting them in a quantitative manner. We also examine systems and devices for preventing electrical hazards.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted into thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the short circuit, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in [link]. A toaster is plugged into a common household electrical outlet. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact, or “short.” As a result, thermal energy can quickly raise the temperature of surrounding materials, melting the insulation and perhaps causing a fire.

The circuit diagram shows a symbol that consists of a sine wave enclosed in a circle. This symbol represents an alternating current (ac) voltage source. In an ac voltage source, the voltage oscillates between a positive and negative maximum amplitude. Up to now, we have been considering direct current (dc) voltage sources, but many of the same concepts are applicable to ac circuits.



A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P = V^2/r$, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.

Another serious thermal hazard occurs when wires supplying power to an appliance are overloaded. Electrical wires and appliances are often rated for the maximum current they can safely handle. The term "overloaded" refers to a condition where the current exceeds the rated maximum current. As current flows through a wire, the power dissipated in the supply wires is $P = I^2 R_{\text{W}}$, where R_{W} is the resistance of the wires and I is the current flowing through the wires. If either I or R_{W} is too large, the wires overheat. Fuses and circuit breakers are used to limit excessive currents.

Shock Hazards

Electric shock is the physiological reaction or injury caused by an external electric current passing through the body. The effect of an electric shock can be negative or positive. When a current with a magnitude above 300 mA passes through the heart, death may occur. Most electrical shock fatalities occur because a current causes ventricular fibrillation, a massively irregular and often fatal, beating of the heart. On the other hand, a heart attack victim, whose heart is in fibrillation, can be saved by an electric shock from a defibrillator.

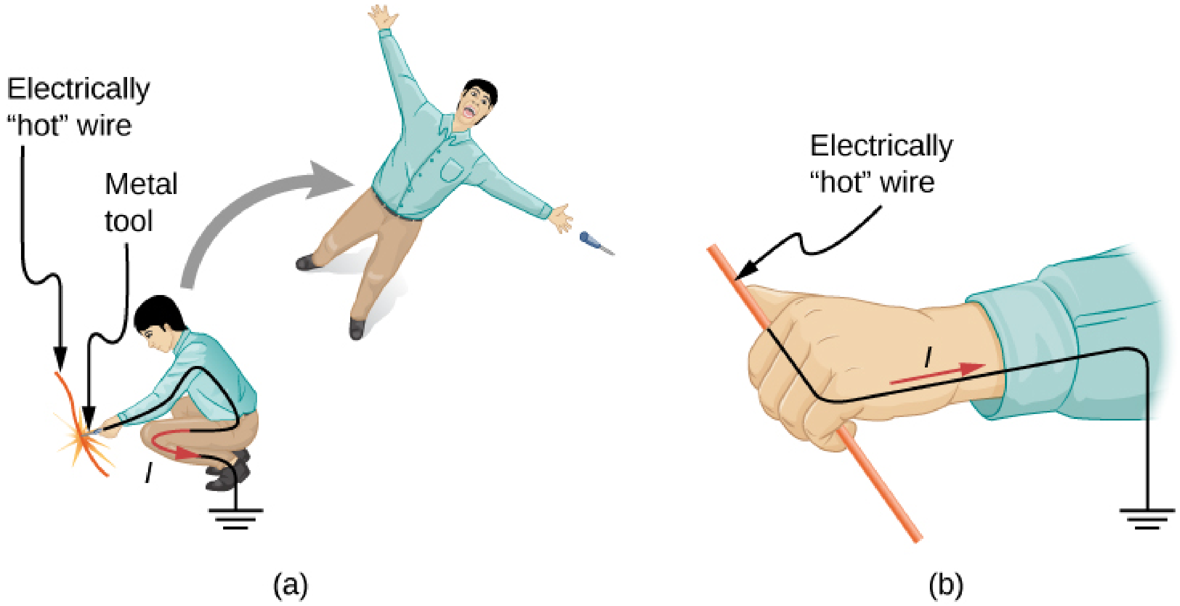
The effects of an undesirable electric shock can vary in severity: a slight sensation at the point of contact, pain, loss of voluntary muscle control, difficulty breathing, heart fibrillation, and possibly death. The loss of voluntary muscle control can cause the victim to not be able to let go of the source of the current.

The major factors upon which the severity of the effects of electrical shock depend are

1. The amount of current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current ($f = 0$ for dc)

Our bodies are relatively good electric conductors due to the body's water content. A dangerous condition occurs when the body is in contact with a voltage source and "ground." The term "ground" refers to a large sink or source of electrons, for example, the earth (thus, the name). When there is a direct path to ground, large currents will pass through the parts of the body with the lowest resistance and a direct path to ground. A safety precaution used by many professions is the wearing of insulated shoes. Insulated shoes prohibit a pathway to ground for electrons through the feet by providing a large resistance. Whenever working with high-power tools, or any electric circuit, ensure that you do not provide a pathway for current flow (especially across the heart). A common safety precaution is to work with one hand, reducing the possibility of providing a current path through the heart.

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 5–30 mA and above, the current can stimulate sustained muscular contractions, much as regular nerve impulses do ([link]). Very large currents (above 300 mA) cause the heart and diaphragm of the lung to contract for the duration of the shock. Both the heart and respiration stop. Both often return to normal following the shock.



An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Current is the major factor determining shock severity. A larger voltage is more hazardous, but since $I = V/R$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance of about $200\text{ k}\Omega$. If he comes into contact with 120-V ac, a current

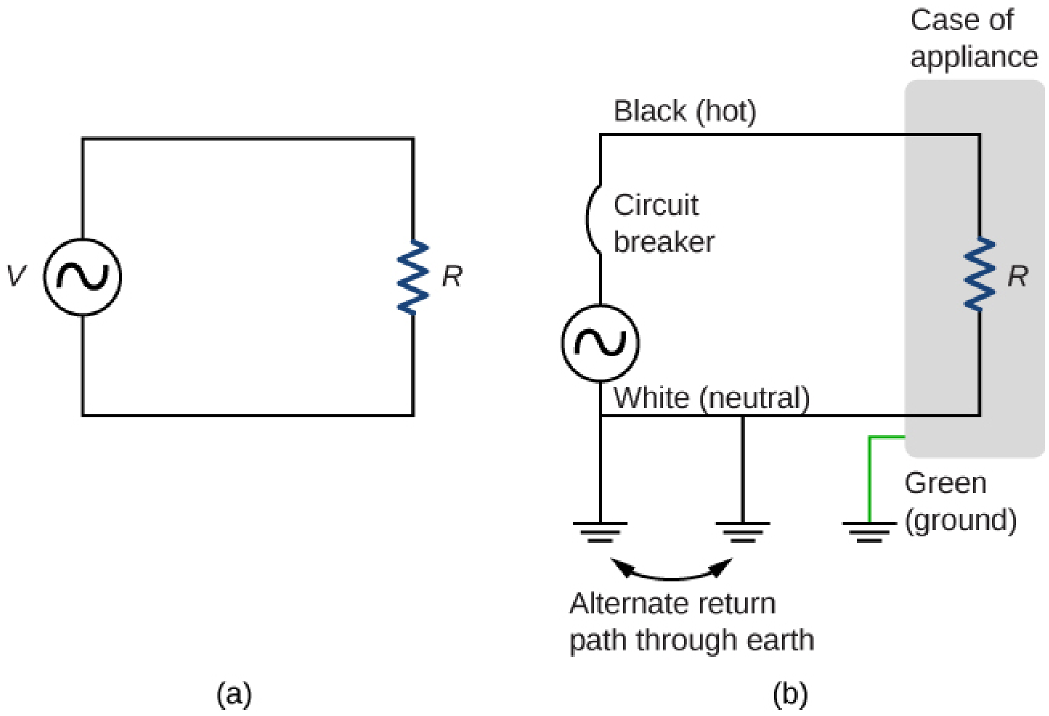
Equation:

$$I = (120\text{ V})/(200\text{ k}\Omega) = 0.6\text{ mA}$$

passes harmlessly through him. The same person soaking wet may have a resistance of $10.0\text{ k}\Omega$ and the same 120 V will produce a current of 12 mA—above the "can't let go" threshold and potentially dangerous.

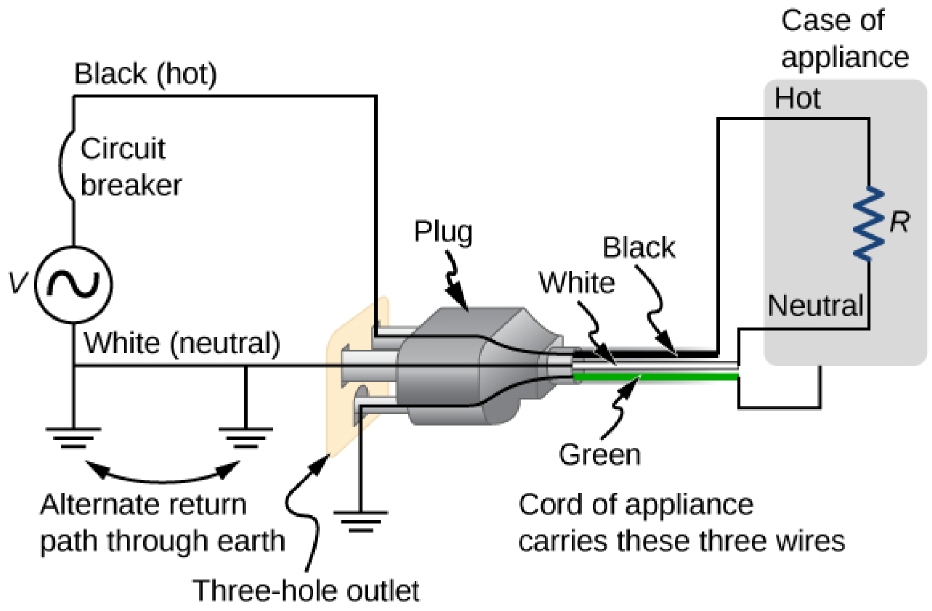
Electrical Safety: Systems and Devices

[link](a) shows the schematic for a simple ac circuit with no safety features. This is not how power is distributed in practice. Modern household and industrial wiring requires the **three-wire system**, shown schematically in part (b), which has several safety features, with live, neutral, and ground wires. First is the familiar circuit breaker (or fuse) to prevent thermal overload. Second is a protective case around the appliance, such as a toaster or refrigerator. The case's safety feature is that it prevents a person from touching exposed wires and coming into electrical contact with the circuit, helping prevent shocks.



(a) Schematic of a simple ac circuit with a voltage source and a single appliance represented by the resistance R. There are no safety features in this circuit. (b) The three-wire system connects the neutral wire to ground at the voltage source and user location, forcing it to be at zero volts and supplying an alternative return path for the current through ground. Also grounded to zero volts is the case of the appliance. A circuit breaker or fuse protects against thermal overload and is in series on the active (live/hot) wire.

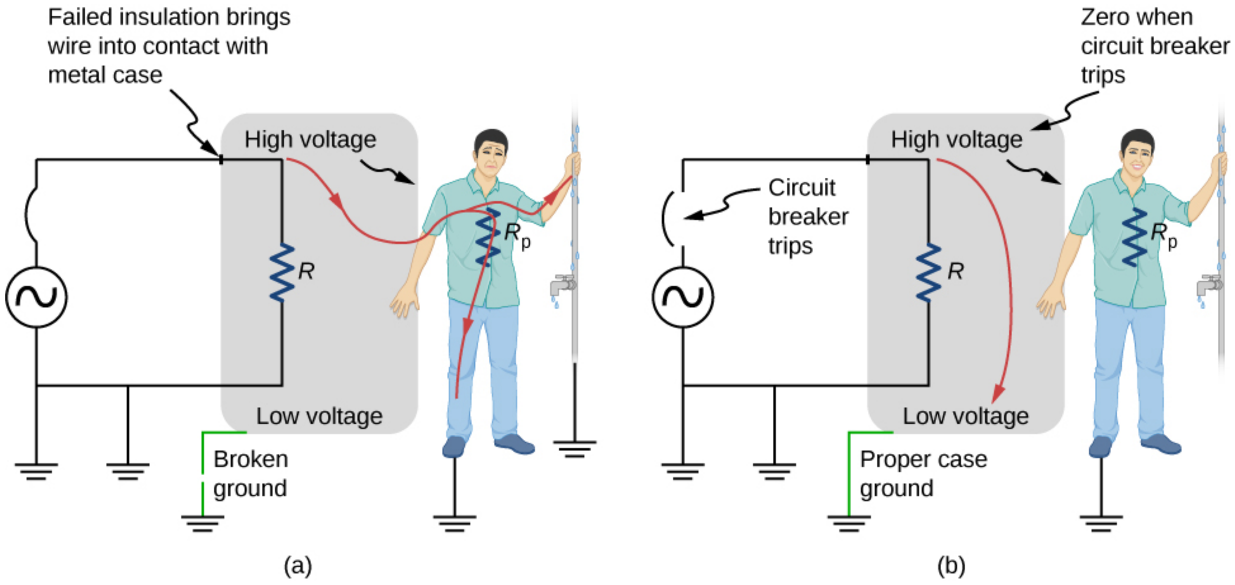
There are three connections to ground shown in [link](b). Recall that a ground connection is a low-resistance path directly to ground. The two ground connections on the neutral wire force it to be at zero volts relative to ground, giving the wire its name. This wire is therefore safe to touch even if its insulation, usually white, is missing. The neutral wire is the return path for the current to follow to complete the circuit. Furthermore, the two ground connections supply an alternative path through ground (a good conductor) to complete the circuit. The ground connection closest to the power source could be at the generating plant, whereas the other is at the user's location. The third ground is to the case of the appliance, through the green ground wire, forcing the case, too, to be at zero volts. The live or hot wire (hereafter referred to as "live/hot") supplies voltage and current to operate the appliance. [link] shows a more pictorial version of how the three-wire system is connected through a three-prong plug to an appliance.


Case of appliance
Black (hot)
Hot
Circuit breaker
R
Plug
Black
V
White
Neutral
White (neutral)
Green
Alternate return path through earth
Cord of appliance carries these three wires
Three-hole outlet

The standard three-prong plug can only be inserted in one way, to ensure proper function of the three-wire system.

Insulating plastic is color-coded to identify live/hot, neutral, and ground wires, but these codes vary around the world. It is essential to determine the color code in your region. Striped coatings are sometimes used for the benefit of those who are colorblind.

Grounding the case solves more than one problem. The simplest problem is worn insulation on the live/hot wire that allows it to contact the case, as shown in [link]. Lacking a ground connection, a severe shock is possible. This is particularly dangerous in the kitchen, where a good connection to ground is available through water on the floor or a water faucet. With the ground connection intact, the circuit breaker will trip, forcing repair of the appliance.


Failed insulation brings wire into contact with metal case
High voltage
R
R_p
Low voltage
Broken ground
(a)
Zero when circuit breaker trips
High voltage
Circuit breaker trips
R
R_p
Low voltage
Proper case ground
(b)

Worn insulation allows the live/hot wire to come into direct contact with the metal case of this appliance. (a) The ground connection being broken, the person is severely shocked. The appliance may operate normally in this situation. (b) With a proper ground, the circuit breaker trips, forcing repair of the appliance.

A ground fault circuit interrupter (GFCI) is a safety device found in updated kitchen and bathroom wiring that works based on electromagnetic induction. GFCIs compare the currents in the live/hot and neutral wires. When live/hot and neutral currents are not equal, it is almost always because current in the neutral is less than in the live/hot wire. Then some of the current, called a leakage current, is returning to the voltage source by a path other than through the neutral wire. It is assumed that this path presents a hazard. GFCIs are usually set to interrupt the circuit if the leakage current is greater than 5 mA, the accepted maximum harmless shock. Even if the leakage current goes safely to ground through an intact ground wire, the GFCI will trip, forcing repair of the leakage.

Summary

- The two types of electric hazards are thermal (excessive power) and shock (current through a person). Electrical safety systems and devices are employed to prevent thermal and shock hazards.
- Shock severity is determined by current, path, duration, and ac frequency.
- Circuit breakers and fuses interrupt excessive currents to prevent thermal hazards.
- The three-wire system guards against thermal and shock hazards, utilizing live/hot, neutral, and ground wires, and grounding the neutral wire and case of the appliance.
- A ground fault circuit interrupter (GFCI) prevents shock by detecting the loss of current to unintentional paths.

Key Equations

Terminal voltage of a single voltage source	$V_{\text{terminal}} = \varepsilon - Ir_{\text{eq}}$
Equivalent resistance of a series circuit	$R_{\text{eq}} = R_1 + R_2 + R_3 + \cdots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i$
Equivalent resistance of a parallel circuit	$R_{\text{eq}} = \left(\frac{1}{R_1} + \frac{1}{R2} + \cdots + \frac{1}{R_N}\right)^{-1} = \left(\sum_{i=1}^{N} \frac{1}{R_i}\right)^{-1}$
Junction rule	$\sum I_{\text{in}} = \sum I_{\text{out}}$
Loop rule	$\sum V = 0$
Terminal voltage of *N* voltage sources in series	$V_{\text{terminal}} = \sum_{i=1}^{N} \varepsilon_i - I\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} \varepsilon_i - Ir_{\text{eq}}$
Terminal voltage of *N* voltage sources in parallel	$V_{\text{terminal}} = \varepsilon - I\sum_{i=1}^{N}\left(\frac{1}{r_i}\right)^{-1} = \varepsilon - Ir_{\text{eq}}$
Charge on a charging capacitor	$q\left(t\right) = C\varepsilon\left(1 - e^{-\frac{t}{RC}}\right) = Q\left(1 - e^{-\frac{t}{\tau}}\right)$
Time constant	$\tau = RC$
Current during charging of a capacitor	$I = \frac{\varepsilon}{R} e^{-\frac{t}{RC}} = I_o e^{-\frac{t}{RC}}$
Charge on a discharging capacitor	$q\left(t\right) = Qe^{-\frac{t}{\tau}}$
Current during discharging of a capacitor	$I\left(t\right) = -\frac{Q}{RC} e^{-\frac{t}{\tau}}$

Conceptual Questions

Exercise:

Problem: Why isn’t a short circuit necessarily a shock hazard?

Exercise:

Problem:

We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why?

Solution:

Not only might water drip into the switch and cause a shock, but also the resistance of your body is lower when you are wet.

Problems

Exercise:

Problem:

(a) How much power is dissipated in a short circuit of 240-V ac through a resistance of $0.250\ \Omega$? (b) What current flows?

Exercise:

Problem:

What voltage is involved in a 1.44-kW short circuit through a $0.100\text{-}\Omega$ resistance?

Solution:

12.0 V

Exercise:

Problem:

Find the current through a person and identify the likely effect on her if she touches a 120-V ac source: (a) if she is standing on a rubber mat and offers a total resistance of $300\ \text{k}\Omega$; (b) if she is standing barefoot on wet grass and has a resistance of only $4000\ \text{k}\Omega$.

Exercise:

Problem:

While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of $4000\ \Omega$. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

Solution:

400 V

Exercise:

Problem:

A man foolishly tries to fish a burning piece of bread from a toaster with a metal butter knife and comes into contact with 120-V ac. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

Exercise:

Problem:

(a) During surgery, a current as small as $20.0\ \mu\text{A}$ applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is $300\ \Omega$, what is the smallest voltage that poses this danger? (b) Does your answer imply that special electrical safety precautions are needed?

Solution:

a. 6.00 mV; b. It would not be necessary to take extra precautions regarding the power coming from the wall. However, it is possible to generate voltages of approximately this value from static charge built up on gloves, for instance, so some precautions are necessary.

Exercise:

Problem:

(a) What is the resistance of a 220-V ac short circuit that generates a peak power of 96.8 kW? (b) What would the average power be if the voltage were 120 V ac?

Exercise:

Problem:

A heart defibrillator passes 10.0 A through a patient’s torso for 5.00 ms in an attempt to restore normal beating. (a) How much charge passed? (b) What voltage was applied if 500 J of energy was dissipated? (c) What was the path’s resistance? (d) Find the temperature increase caused in the 8.00 kg of affected tissue.

Solution:

a. $5.00\ \times\ 10^{-2}\ \text{C}$; b. 10.0 kV; c. $1.00\ \text{k}\Omega$; d. $1.79\ \times\ 10^{-2}\ {}^{\circ}\text{C}$

Exercise:

Problem:

A short circuit in a 120-V appliance cord has a $0.500\text{-}\Omega$ resistance. Calculate the temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is $0.200 \text{ cal/g} \cdot {}^{\circ}\text{C}$ and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

Additional Problems

Exercise:

Problem:

A circuit contains a D cell battery, a switch, a $20\text{-}\Omega$ resistor, and four 20-mF capacitors connected in series. (a) What is the equivalent capacitance of the circuit? (b) What is the *RC* time constant? (c) How long before the current decreases to 50% of the initial value once the switch is closed?

Solution:

a. $C_{\text{eq}} = 5.00 \text{ mF}$; b. $\tau = 0.1$ s; c. 0.069 s

Exercise:

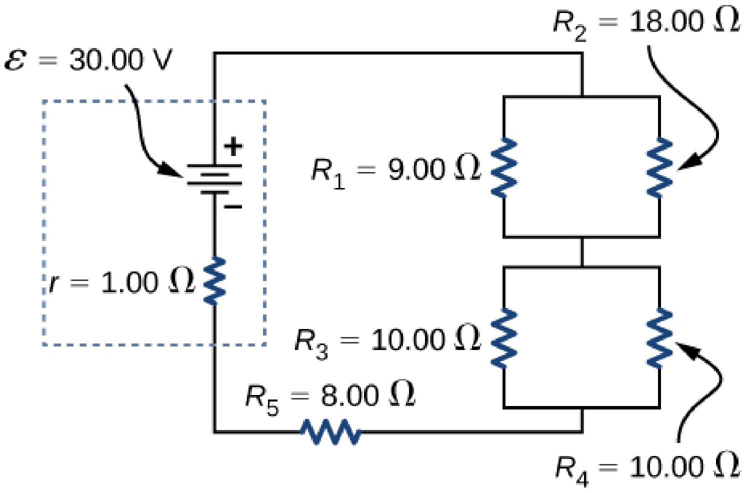
Problem:

A circuit contains a D-cell battery, a switch, a $20\text{-}\Omega$ resistor, and three 20-mF capacitors. The capacitors are connected in parallel, and the parallel connection of capacitors are connected in series with the switch, the resistor and the battery. (a) What is the equivalent capacitance of the circuit? (b) What is the *RC* time constant? (c) How long before the current decreases to 50% of the initial value once the switch is closed?

Exercise:

Problem:

Consider the circuit below. The battery has an emf of $\varepsilon = 30.00 \text{ V}$ and an internal resistance of $r = 1.00 \,\Omega$. (a) Find the equivalent resistance of the circuit and the current out of the battery. (b) Find the current through each resistor. (c) Find the potential drop across each resistor. (d) Find the power dissipated by each resistor. (e) Find the total power supplied by the batteries.



Solution:

a. $R_{\text{eq}} = 20.00\ \Omega$;
b. $I_r = 1.50\ \text{A}, I_1 = 1.00\ \text{A}, I_2 = 0.50\ \text{A}, I_3 = 0.75\ \text{A}, I_4 = 0.75\ \text{A}, I_5 = 1.50\ \text{A}$;
c. $V_r = 1.50\ \text{V}, V_1 = 9.00\ \text{V}, V_2 = 9.00\ \text{V}, V_3 = 7.50\ \text{V}, V_4 = 7.50\ \text{V}, V_5 = 12.00\ \text{V}$;
d.
$P_r = 2.25\ \text{W}, P_1 = 9.00\ \text{W}, P_2 = 4.50\ \text{W}, P_3 = 5.625\ \text{W}, P_4 = 5.625\ \text{W}, P_5 = 18.00\ \text{W}$
;
e. $P = 45.00\ \text{W}$

Exercise:

Problem:

A homemade capacitor is constructed of 2 sheets of aluminum foil with an area of 2.00 square meters, separated by paper, 0.05 mm thick, of the same area and a dielectric constant of 3.7. The homemade capacitor is connected in series with a 100.00-Ω resistor, a switch, and a 6.00-V voltage source. (a) What is the *RC* time constant of the circuit? (b) What is the initial current through the circuit, when the switch is closed? (c) How long does it take the current to reach one third of its initial value?

Exercise:

Problem:

A student makes a homemade resistor from a graphite pencil 5.00 cm long, where the graphite is 0.05 mm in diameter. The resistivity of the graphite is $\rho = 1.38 \times 10^{-5}\ \Omega/\text{m}$. The homemade resistor is place in series with a switch, a 10.00-mF capacitor and a 0.50-V power source. (a) What is the *RC* time constant of the circuit? (b) What is the potential drop across the pencil 1.00 s after the switch is closed?

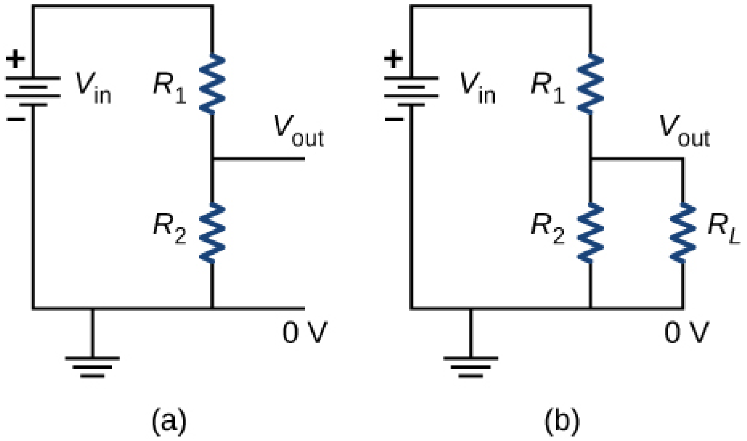
Solution:

a. $\tau = \left(1.38\ \times\ 10^{-5}\Omega\text{m}\left(\frac{5.00\times10^{-2}\,\text{m}}{3.14\left(\frac{0.05\times10^{-3}}{2}\right)^2}\right)\right)10\ \times\ 10^{-3}\,\text{F} = 3.52\,\text{s}$; b. $V = 0.017\,\text{A}\left(e^{-\frac{1.00\,\text{s}}{3.52\,\text{s}}}\right)351.59\,\Omega = 0.122\,\text{V}$

Exercise:

Problem:

The rather simple circuit shown below is known as a voltage divider. The symbol consisting of three horizontal lines is represents “ground” and can be defined as the point where the potential is zero. The voltage divider is widely used in circuits and a single voltage source can be used to provide reduced voltage to a load resistor as shown in the second part of the figure. (a) What is the output voltage V_{out} of circuit (a) in terms of $R_1, R_2,$ and V_{in}? (b) What is the output voltage V_{out} of circuit (b) in terms of $R_1, R_2, R_L,$ and V_{in}?



Exercise:

Problem:

Three 300-Ω resistors are connect in series with an AAA battery with a rating of 3 AmpHours. (a) How long can the battery supply the resistors with power? (b) If the resistors are connected in parallel, how long can the battery last?

Solution:

a. $t = \frac{3\,\text{A·h}}{\frac{1.5\,\text{V}}{900\,\Omega}} = 1800\,\text{h}$; b. $t = \frac{3\,\text{A·h}}{\frac{1.5\,\text{V}}{100\,\Omega}} = 200\,\text{h}$

Exercise:

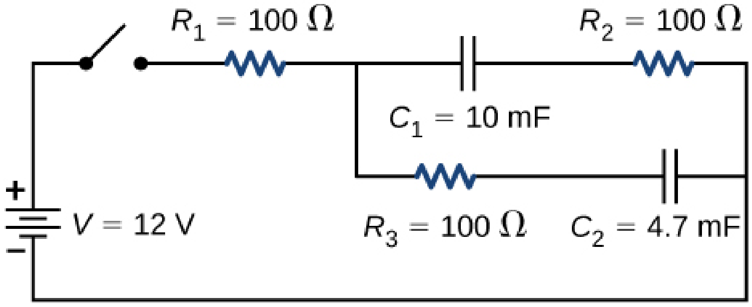
Problem:

Consider a circuit that consists of a real battery with an emf ε and an internal resistance of r connected to a variable resistor R. (a) In order for the terminal voltage of the battery to be equal to the emf of the battery, what should the resistance of the variable resistor be adjusted to? (b) In order to get the maximum current from the battery, what should the resistance of the variable resistor be adjusted to? (c) In order for the maximum power output of the battery to be reached, what should the resistance of the variable resistor be set to?

Exercise:

Problem:

Consider the circuit shown below. What is the energy stored in each capacitor after the switch has been closed for a very long time?



Solution:

$$U_1 = C_1V_1^2 = 0.72\text{J}, \;\; U_2 = C_2V_2^2 = 0.338 \text{ J}$$

Exercise:

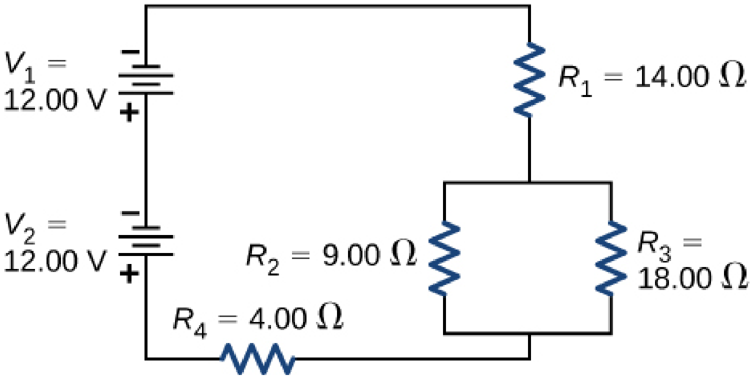
Problem:

Consider a circuit consisting of a battery with an emf ε and an internal resistance of r connected in series with a resistor R and a capacitor C. Show that the total energy supplied by the battery while charging the battery is equal to $\varepsilon^2 C$.

Exercise:

Problem:

Consider the circuit shown below. The terminal voltages of the batteries are shown. (a) Find the equivalent resistance of the circuit and the current out of the battery. (b) Find the current through each resistor. (c) Find the potential drop across each resistor. (d) Find the power dissipated by each resistor. (e) Find the total power supplied by the batteries.


$V_1 =$ 12.00 V
$V_2 =$ 12.00 V
$R_1 = 14.00\ \Omega$
$R_2 = 9.00\ \Omega$
$R_3 = 18.00\ \Omega$
$R_4 = 4.00\ \Omega$

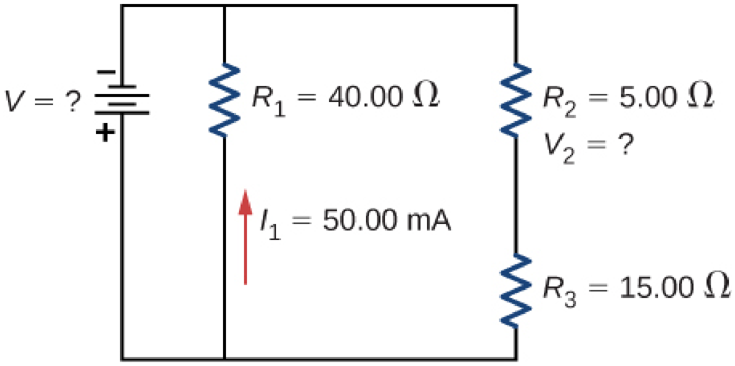
Solution:

a. $R_{\text{eq}} = 24.00\ \Omega$; b. $I_1 = 1.00\ \text{A}, I_2 = 0.67\ \text{A}, I_3 = 0.33\ \text{A}, I_4 = 1.00\ \text{A}$;
c. $V_1 = 14.00\ \text{V}, V_2 = 6.00\ \text{V}, V_3 = 6.00\ \text{V}, V_4 = 4.00\ \text{V}$;
d. $P_1 = 14.00\ \text{W}, P_2 = 4.04\ \text{W}, P_3 = 1.96\ \text{W}, P_4 = 4.00\ \text{W}$; e. $P = 24.00\ \text{W}$

Exercise:

Problem:

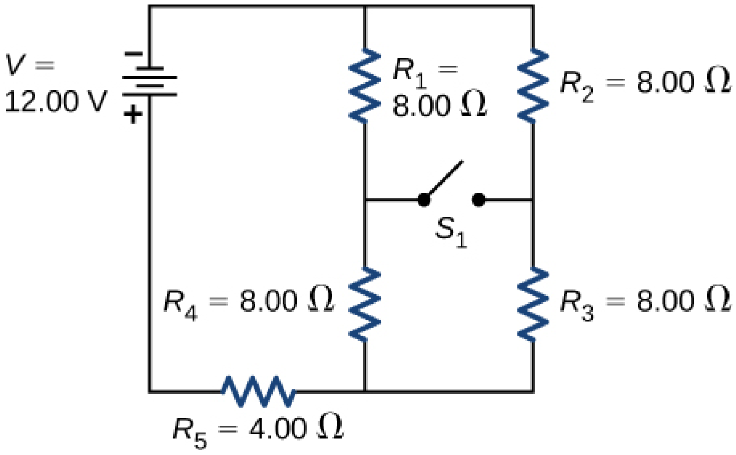
Consider the circuit shown below. (a) What is the terminal voltage of the battery? (b) What is the potential drop across resistor R_2?


V = ?
$R_1 = 40.00\ \Omega$
$I_1 = 50.00$ mA
$R_2 = 5.00\ \Omega$
$V_2 = ?$
$R_3 = 15.00\ \Omega$

Exercise:

Problem:

Consider the circuit shown below. (a)Determine the equivalent resistance and the current from the battery with switch S_1 open. (b) Determine the equivalent resistance and the current from the battery with switch S_1 closed.



Solution:

a. $R_{\text{eq}} = 12.00\ \Omega, I = 1.00\ \text{A}$; b. $R_{\text{eq}} = 12.00\ \Omega, I = 1.00\ \text{A}$

Exercise:

Problem:

Two resistors, one having a resistance of $145\ \Omega$, are connected in parallel to produce a total resistance of $150\ \Omega$. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem:

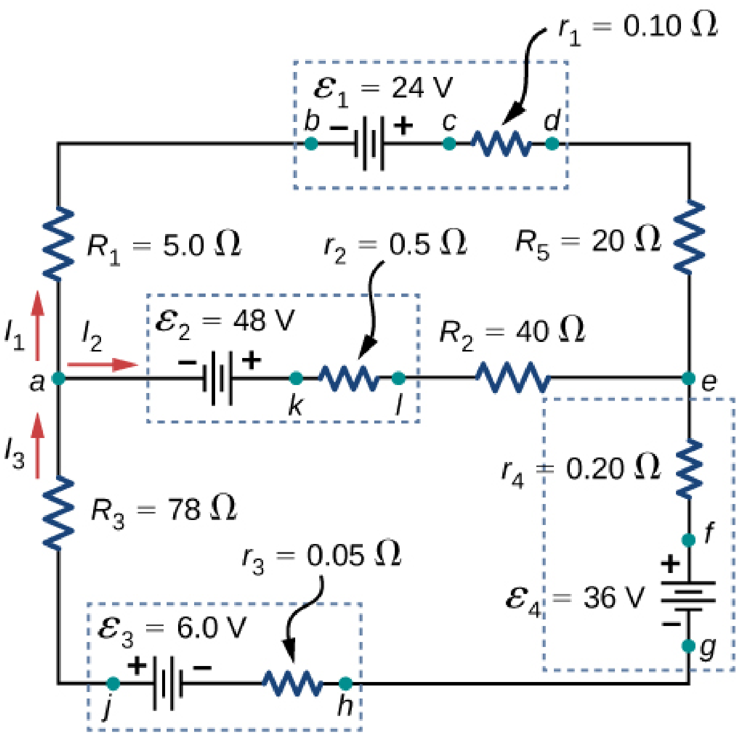
Two resistors, one having a resistance of $900\ \text{k}\Omega$, are connected in series to produce a total resistance of $0.500\ \text{M}\Omega$. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. $-400\ \text{k}\Omega$; b. You cannot have negative resistance. c. The assumption that $R_{\text{eq}} < R_1$ is unreasonable. Series resistance is always greater than any of the individual resistances.

Exercise:

Problem: Apply the junction rule at point *a* shown below.



$r_1 = 0.10\ \Omega$
$\mathcal{E}_1 = 24$ V
b
c
d
$R_1 = 5.0\ \Omega$
$r_2 = 0.5\ \Omega$
$R_5 = 20\ \Omega$
I_1
I_2
$\mathcal{E}_2 = 48$ V
$R_2 = 40\ \Omega$
a
k
l
e
I_3
$r_4 = 0.20\ \Omega$
f
$R_3 = 78\ \Omega$
$r_3 = 0.05\ \Omega$
$\mathcal{E}_4 = 36$ V
g
$\mathcal{E}_3 = 6.0$ V
j
h

Exercise:

Problem: Apply the loop rule to Loop *akledcba* in the preceding problem.

Solution:

$$E_2 - I_2 r_2 - I_2 R_2 + I_1 R_5 + I_1 r_1 - E_1 + I_1 R_1 = 0$$

Exercise:

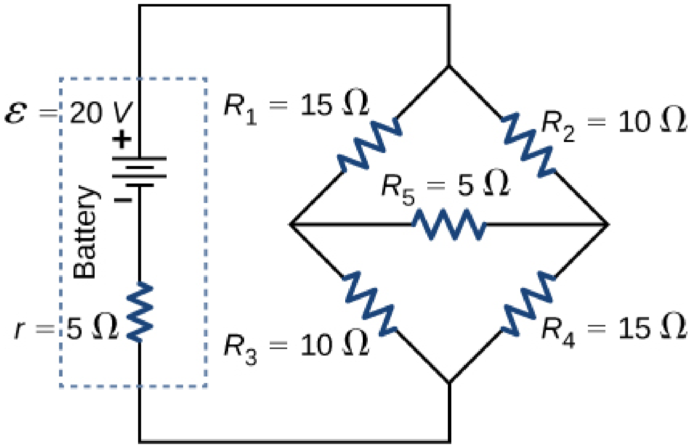
Problem:

Find the currents flowing in the circuit in the preceding problem. Explicitly show how you follow the steps in the Problem-Solving Strategy: Series and Parallel Resistors.

Exercise:

Problem:

Consider the circuit shown below. (a) Find the current through each resistor. (b) Check the calculations by analyzing the power in the circuit.



Solution:

a. $I = 1.17 \text{ A}, I_1 = 0.50 \text{ A}, I_2 = 0.67 \text{ A}, I_3 = 0.67 \text{ A}, I_4 = 0.50 \text{ A}, I_5 = 0.17 \text{ A}$;
b. $P_{\text{output}} = 23.4 \text{ W}, P_{\text{input}} = 23.4 \text{ W}$

Exercise:

Problem:

A flashing lamp in a Christmas earring is based on an *RC* discharge of a capacitor through its resistance. The effective duration of the flash is 0.250 s, during which it produces an average 0.500 W from an average 3.00 V. (a) What energy does it dissipate? (b) How much charge moves through the lamp? (c) Find the capacitance. (d) What is the resistance of the lamp? (Since average values are given for some quantities, the shape of the pulse profile is not needed.)

Exercise:

Problem:

A 160-μF capacitor charged to 450 V is discharged through a $31.2\text{-k}\Omega$ resistor. (a) Find the time constant. (b) Calculate the temperature increase of the resistor, given that its mass is 2.50 g and its specific heat is $1.67 \text{ kJ/kg} \cdot {}^\circ\text{C}$, noting that most of the thermal energy is retained in the short time of the discharge. (c) Calculate the new resistance, assuming it is pure carbon. (d) Does this change in resistance seem significant?

Solution:

a. 4.99 s; b. $3.87\ {}^\circ\text{C}$; c. $3.11 \times 10^4\ \Omega$; d. No, this change does not seem significant. It probably would not be noticed.

Challenge Problems

Exercise:

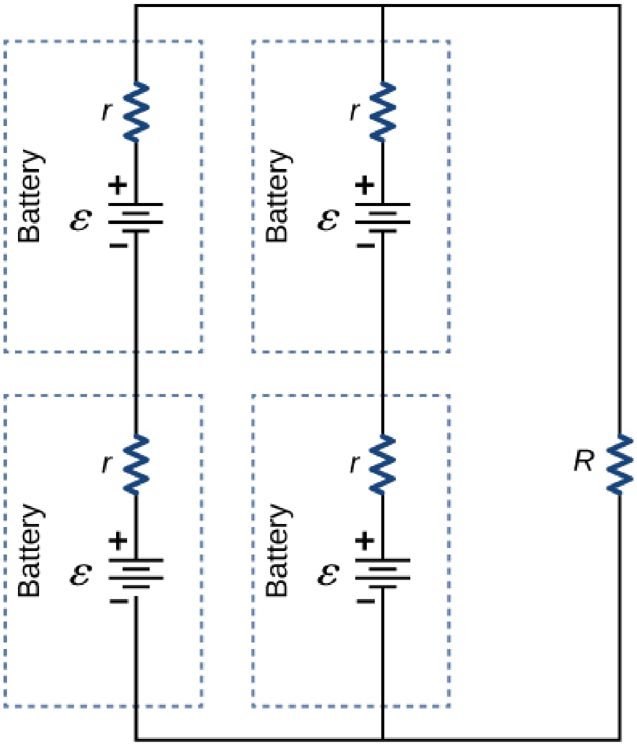
Problem:

Some camera flashes use flash tubes that require a high voltage. They obtain a high voltage by charging capacitors in parallel and then internally changing the connections of the capacitors to place them in series. Consider a circuit that uses four AAA batteries connected in series to charge six 10-mF capacitors through an equivalent resistance of $100\ \Omega$. The connections are then switched internally to place the capacitors in series. The capacitors discharge through a lamp with a resistance of $100\ \Omega$. (a) What is the *RC* time constant and the initial current out of the batteries while they are connected in parallel? (b) How long does it take for the capacitors to charge to 90% of the terminal voltages of the batteries? (c) What is the *RC* time constant and the initial current of the capacitors connected in series assuming it discharges at 90% of full charge? (d) How long does it take the current to decrease to 10% of the initial value?

Exercise:

Problem:

Consider the circuit shown below. Each battery has an emf of 1.50 V and an internal resistance of $1.00\ \Omega$. (a) What is the current through the external resistor, which has a resistance of 10.00 ohms? (b) What is the terminal voltage of each battery?



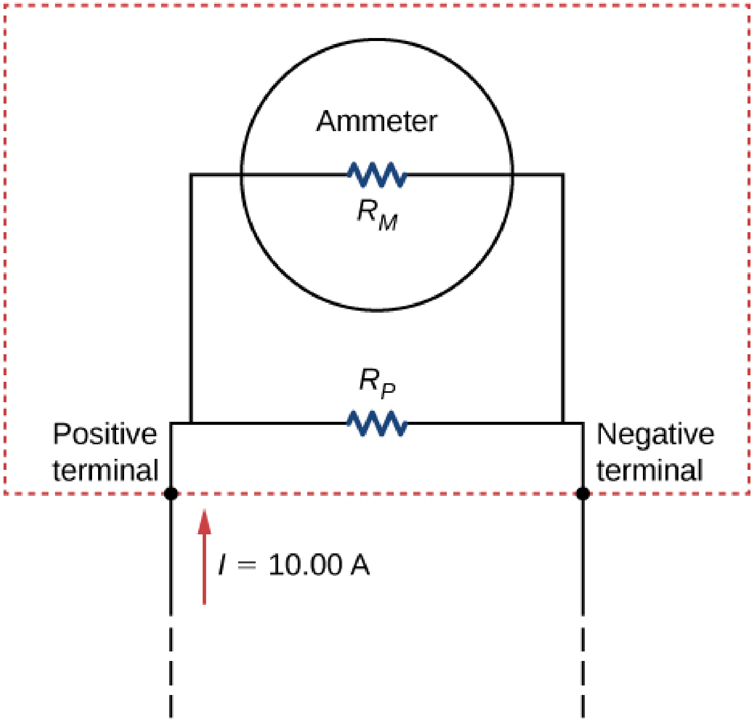
Solution:

a. 0.273 A; b. $V_T = 1.36\ \text{V}$

Exercise:

Problem:

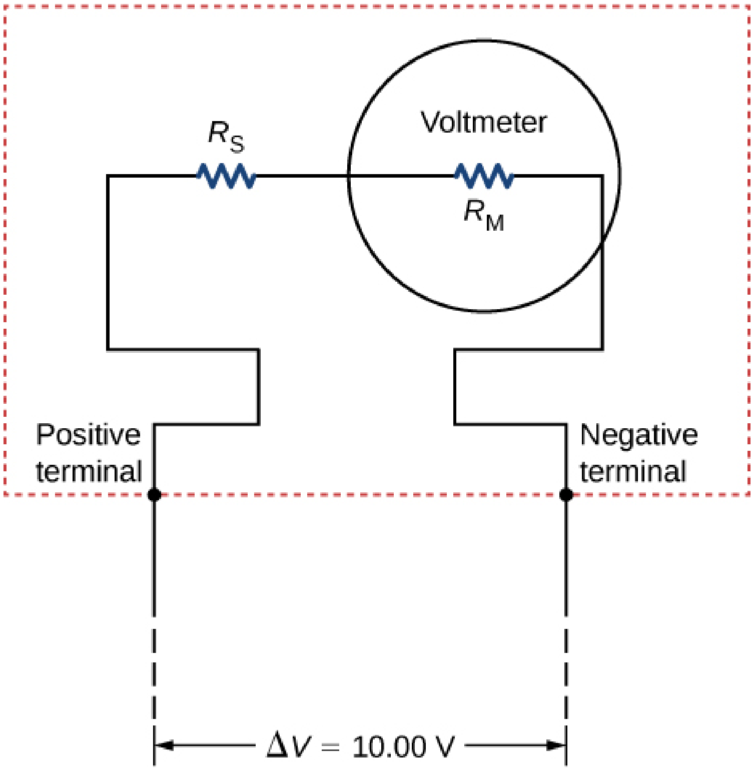
Analog meters use a galvanometer, which essentially consists of a coil of wire with a small resistance and a pointer with a scale attached. When current runs through the coil, the pointer turns; the amount the pointer turns is proportional to the amount of current running through the coil. Galvanometers can be used to make an ammeter if a resistor is placed in parallel with the galvanometer. Consider a galvanometer that has a resistance of $25.00\ \Omega$ and gives a full scale reading when a $50\text{-}\mu\text{A}$ current runs through it. The galvanometer is to be used to make an ammeter that has a full scale reading of 10.00 A, as shown below. Recall that an ammeter is connected in series with the circuit of interest, so all 10 A must run through the meter. (a) What is the current through the parallel resistor in the meter? (b) What is the voltage across the parallel resistor? (c) What is the resistance of the series resistor?



Exercise:

Problem:

Analog meters use a galvanometer, which essentially consists of a coil of wire with a small resistance and a pointer with a scale attached. When current runs through the coil, the point turns; the amount the pointer turns is proportional to the amount of current running through the coil. Galvanometers can be used to make a voltmeter if a resistor is placed in series with the galvanometer. Consider a galvanometer that has a resistance of $25.00\ \Omega$ and gives a full scale reading when a $50\text{-}\mu\text{A}$ current runs through it. The galvanometer is to be used to make an voltmeter that has a full scale reading of 10.00 V, as shown below. Recall that a voltmeter is connected in parallel with the component of interest, so the meter must have a high resistance or it will change the current running through the component. (a) What is the potential drop across the series resistor in the meter? (b) What is the resistance of the parallel resistor?

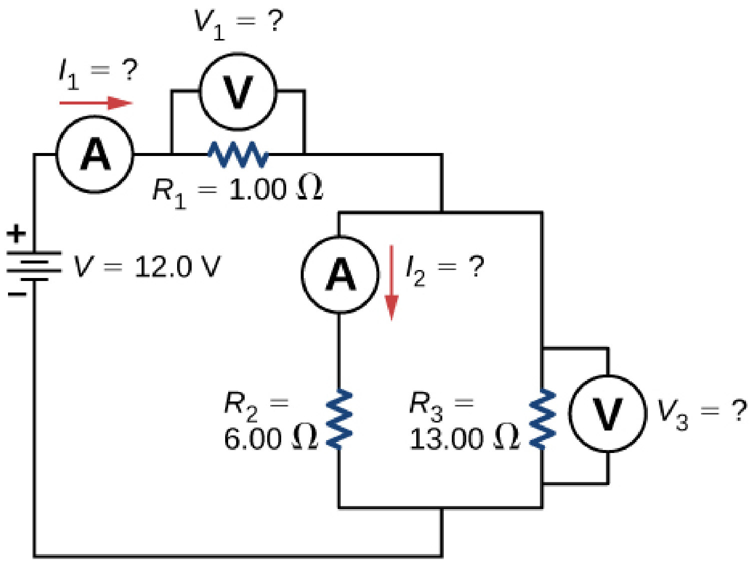


Solution:

a. $V_s = V - I_M R_M = 9.99875\ \text{V}$; b. $R_S = \frac{V_P}{I_M} = 199.975\ \text{k}\Omega$

Exercise:

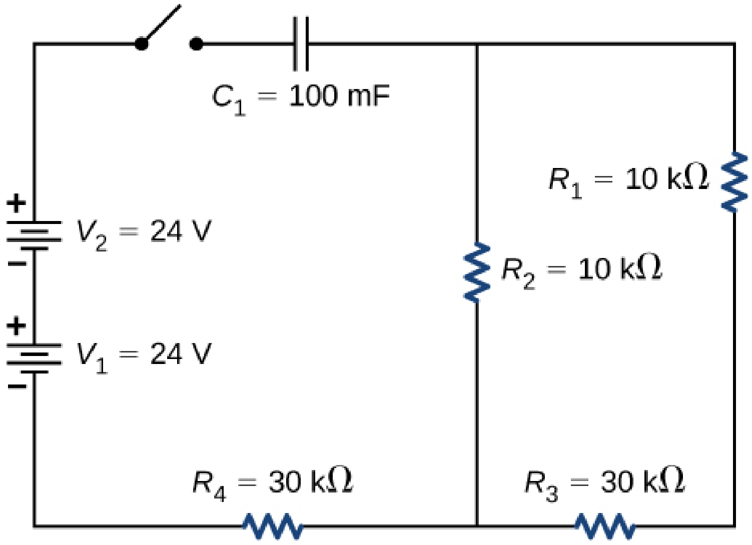
Problem: Consider the circuit shown below. Find $I_1, V_1, I_2,$ and V_3.


V_1 = ?
I_1 = ?
V
A
R_1 = 1.00 Ω
V = 12.0 V
A
I_2 = ?
R_2 = 6.00 Ω
R_3 = 13.00 Ω
V
V_3 = ?

Exercise:

Problem:

Consider the circuit below. (a) What is the *RC* time constant of the circuit? (b) What is the initial current in the circuit once the switch is closed? (c) How much time passes between the instant the switch is closed and the time the current has reached half of the initial current?


C_1 = 100 mF
R_1 = 10 kΩ
V_2 = 24 V
R_2 = 10 kΩ
V_1 = 24 V
R_4 = 30 kΩ
R_3 = 30 kΩ

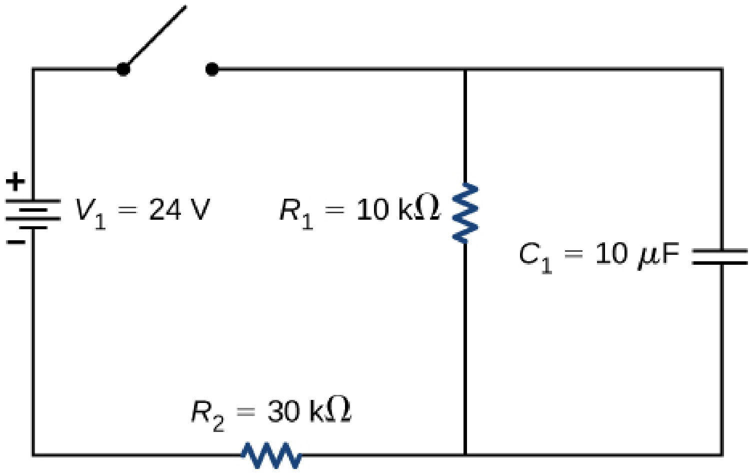
Solution:

a. $\tau = 3800 \text{ s}$; b. 1.26 mA; c. $t = 2633.96 \text{ s}$

Exercise:

Problem:

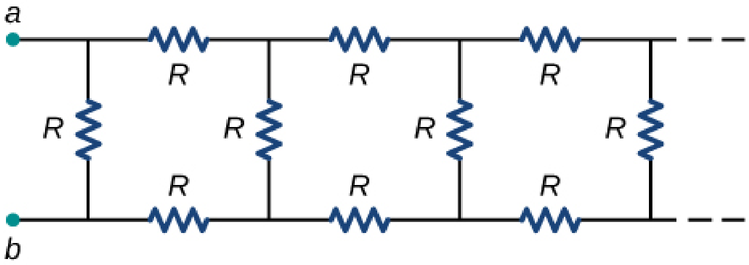
Consider the circuit below. (a) What is the initial current through resistor R_2 when the switch is closed? (b) What is the current through resistor R_2 when the capacitor is fully charged, long after the switch is closed? (c) What happens if the switch is opened after it has been closed for some time? (d) If the switch has been closed for a time period long enough for the capacitor to become fully charged, and then the switch is opened, how long before the current through resistor R_1 reaches half of its initial value?



Exercise:

Problem:

Consider the infinitely long chain of resistors shown below. What is the resistance between terminals *a* and *b*?



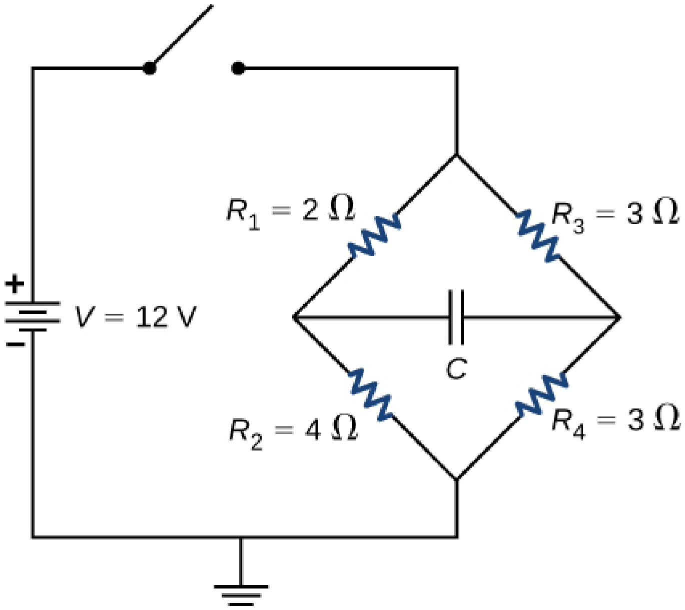
Solution:

$$R_{\text{eq}} = \left(\sqrt{3}–1\right)R$$

Exercise:

Problem:

Consider the circuit below. The capacitor has a capacitance of 10 mF. The switch is closed and after a long time the capacitor is fully charged. (a) What is the current through each resistor a long time after the switch is closed? (b) What is the voltage across each resistor a long time after the switch is closed? (c) What is the voltage across the capacitor a long time after the switch is closed? (d) What is the charge on the capacitor a long time after the switch is closed? (e) The switch is then opened. The capacitor discharges through the resistors. How long from the time before the current drops to one fifth of the initial value?



Exercise:

Problem:

A 120-V immersion heater consists of a coil of wire that is placed in a cup to boil the water. The heater can boil one cup of $20.00\ ^\circ\text{C}$ water in 180.00 seconds. You buy one to use in your dorm room, but you are worried that you will overload the circuit and trip the 15.00-A, 120-V circuit breaker, which supplies your dorm room. In your dorm room, you have four 100.00-W incandescent lamps and a 1500.00-W space heater. (a) What is the power rating of the immersion heater? (b) Will it trip the breaker when everything is turned on? (c) If it you replace the incandescent bulbs with 18.00-W LED, will the breaker trip when everything is turned on?

Solution:

a. $P_{\text{imheater}} = \dfrac{1\text{cup}\left(\frac{0.000237\ \text{m}^3}{\text{cup}}\right)\left(\frac{1000\ \text{kg}}{\text{m}^3}\right)\left(4186\frac{\text{J}}{\text{kg}\ ^\circ\text{C}}\right)(100\ ^\circ\text{C}{-}20\ ^\circ\text{C})}{180.00\ \text{s}} \approx 441\ \text{W};$

b. $I = \frac{441\ \text{W}}{120\ \text{V}} + 4\ \ \frac{100\ \text{W}}{120\ \text{V}}\ \ + \frac{1500\ \text{W}}{120\ \text{V}} = 19.51\ \text{A}$; Yes, the breaker will trip.

c. $I = \frac{441\ \text{W}}{120\ \text{V}} + 4\ \ \frac{18\ \text{W}}{120\ \text{V}}\ \ + \frac{1500\ \text{W}}{120\ \text{V}} = 16.78\ \text{A}$; Yes, the breaker will trip.

Exercise:

Problem:

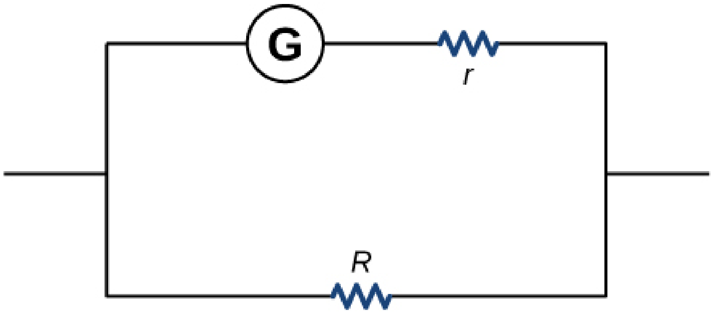
Find the resistance that must be placed in series with a $25.0\text{-}\Omega$ galvanometer having a $50.0\text{-}\mu\text{A}$ sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 3000-V full-scale reading. Include a circuit diagram with your solution.

Exercise:

Problem:

Find the resistance that must be placed in parallel with a $60.0\text{-}\Omega$ galvanometer having a 1.00-mA sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 25.0-A full-scale reading. Include a circuit diagram with your solution.

Solution:



,
$2.40 \times 10^{-3}\,\Omega$

Glossary

shock hazard
hazard in which an electric current passes through a person

thermal hazard
hazard in which an excessive electric current causes undesired thermal effects

three-wire system
wiring system used at present for safety reasons, with live, neutral, and ground wires

Introduction class="introduction"

An industrial electromagnet is capable of lifting thousands of pounds of metallic waste. (credit: modification of work by “BedfordAl”/Flickr)

☎ 0782
87637
J.K. STURGE LTD.
J.K. STUR

For the past few chapters, we have been studying electrostatic forces and fields, which are caused by electric charges at rest. These electric fields can move other free charges, such as producing a current in a circuit; however, the electrostatic forces and fields themselves come from other static charges. In this chapter, we see that when an electric charge moves, it generates other forces and fields. These additional forces and fields are what we commonly call magnetism.

Before we examine the origins of magnetism, we first describe what it is and how magnetic fields behave. Once we are more familiar with magnetic effects, we can explain how they arise from the behavior of atoms and molecules, and how magnetism is related to electricity. The connection between electricity and magnetism is fascinating from a theoretical point of view, but it is also immensely practical, as shown by an industrial electromagnet that can lift thousands of pounds of metal.

Magnetism and Its Historical Discoveries

By the end of this section, you will be able to:

- Explain attraction and repulsion by magnets
- Describe the historical and contemporary applications of magnetism

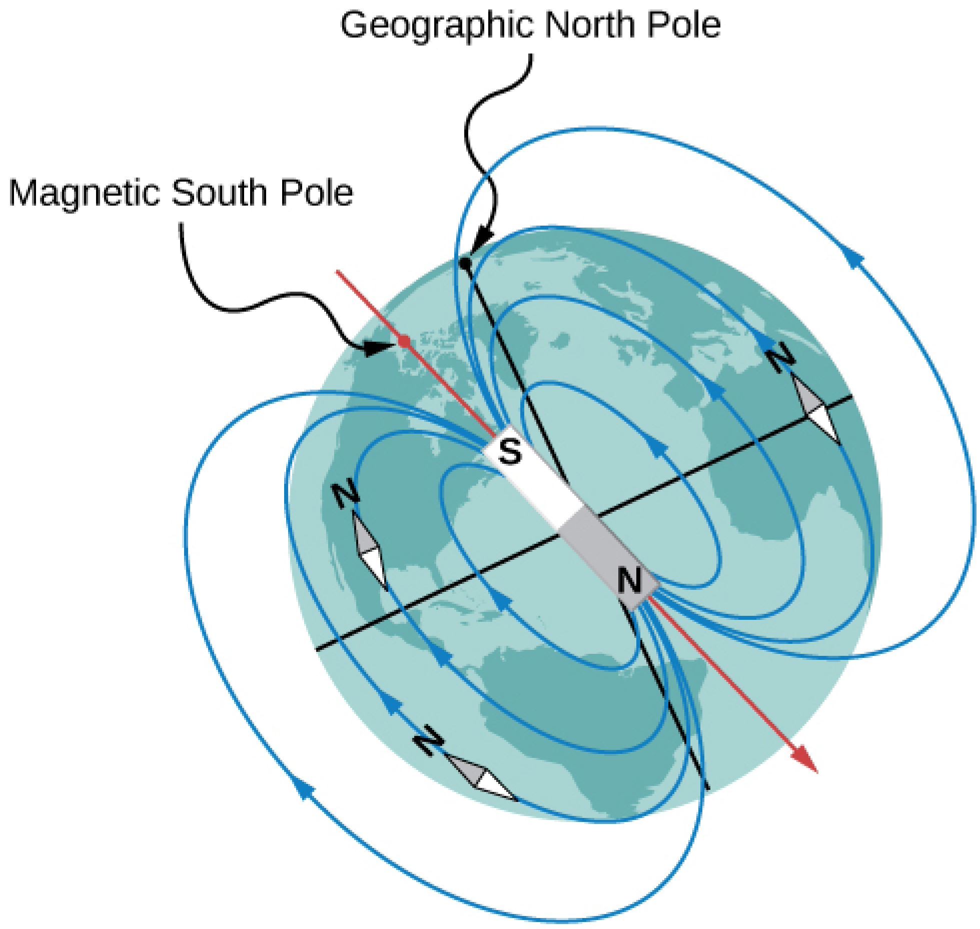
Magnetism has been known since the time of the ancient Greeks, but it has always been a bit mysterious. You can see electricity in the flash of a lightning bolt, but when a compass needle points to magnetic north, you can't see any force causing it to rotate. People learned about magnetic properties gradually, over many years, before several physicists of the nineteenth century connected magnetism with electricity. In this section, we review the basic ideas of magnetism and describe how they fit into the picture of a magnetic field.

Brief History of Magnetism

Magnets are commonly found in everyday objects, such as toys, hangers, elevators, doorbells, and computer devices. Experimentation on these magnets shows that all magnets have two poles: One is labeled north (N) and the other is labeled south (S). Magnetic poles repel if they are alike (both N or both S), they attract if they are opposite (one N and the other S), and both poles of a magnet attract unmagnetized pieces of iron. An important point to note here is that you cannot isolate an individual magnetic pole. Every piece of a magnet, no matter how small, which contains a north pole must also contain a south pole.

Note:
Visit this website for an interactive demonstration of magnetic north and south poles.

An example of a magnet is a compass needle. It is simply a thin bar magnet suspended at its center, so it is free to rotate in a horizontal plane. Earth itself also acts like a very large bar magnet, with its south-seeking pole near the geographic North Pole ([link]). The north pole of a compass is attracted toward Earth’s geographic North Pole because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole. Confusion arises because the geographic term “North Pole” has come to be used (incorrectly) for the magnetic pole that is near the North Pole. Thus, “**north magnetic pole**” is actually a misnomer—it should be called the **south magnetic pole**. [Note that the orientation of Earth’s magnetic field is not permanent but changes (“flips”) after long time intervals. Eventually, Earth’s north magnetic pole may be located near its geographic North Pole.]



The north pole of a compass needle points toward the south pole of a magnet, which is how today’s magnetic field is oriented from inside Earth. It also points toward Earth’s geographic North Pole because the geographic North Pole is near the magnetic south pole.

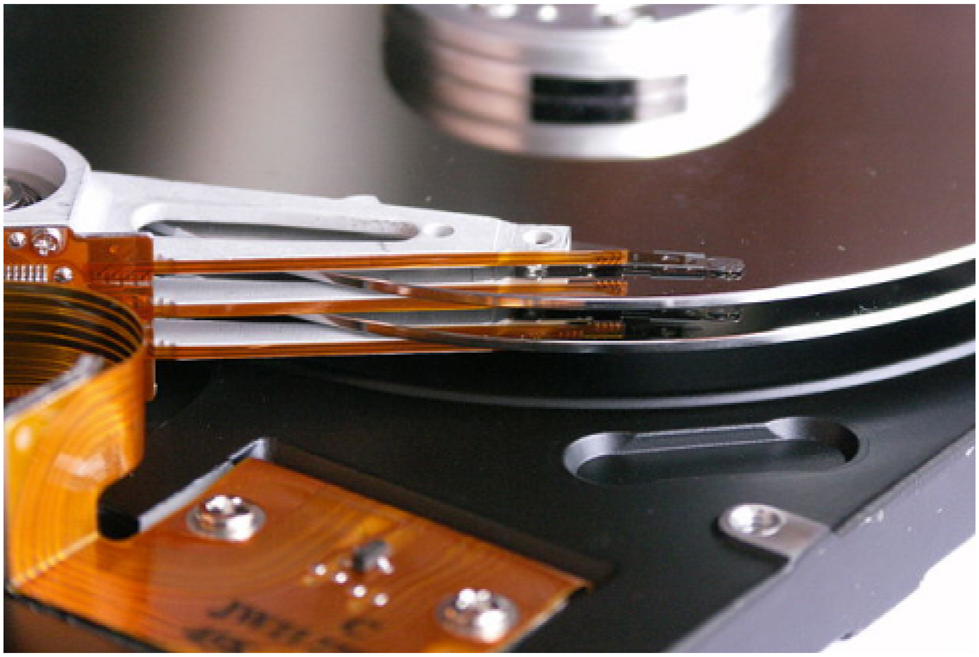
Back in 1819, the Danish physicist Hans Oersted was performing a lecture demonstration for some students and noticed that a compass needle moved whenever current flowed in a nearby wire. Further investigation of this phenomenon convinced Oersted that an electric current could somehow cause a magnetic force. He reported this finding to an 1820 meeting of the French Academy of Science.

Soon after this report, Oersted’s investigations were repeated and expanded upon by other scientists. Among those whose work was especially important were Jean-Baptiste Biot and Felix Savart, who investigated the forces exerted on magnets by currents; André Marie Ampère, who studied the forces exerted by one current on another; François Arago, who found that iron could be magnetized by a current; and Humphry Davy, who discovered that a magnet exerts a force on a wire carrying an electric current. Within 10 years of Oersted’s discovery, Michael Faraday found that the relative motion of a magnet and a metallic wire induced current in the wire. This finding showed not only that a current has a magnetic effect, but that a magnet can generate electric current. You will see later that the names of Biot, Savart, Ampère, and Faraday are linked to some of the fundamental laws of electromagnetism.

The evidence from these various experiments led Ampère to propose that electric current is the source of all magnetic phenomena. To explain permanent magnets, he suggested that matter contains microscopic current loops that are somehow aligned when a material is magnetized. Today, we know that permanent magnets are actually created by the alignment of spinning electrons, a situation quite similar to that proposed by Ampère. This model of permanent magnets was developed by Ampère almost a century before the atomic nature of matter was understood. (For a full quantum mechanical treatment of magnetic spins, see Quantum Mechanics and Atomic Structure.)

Contemporary Applications of Magnetism

Today, magnetism plays many important roles in our lives. Physicists' understanding of magnetism has enabled the development of technologies that affect both individuals and society. The electronic tablet in your purse or backpack, for example, wouldn't have been possible without the applications of magnetism and electricity on a small scale ([link]). Weak changes in a magnetic field in a thin film of iron and chromium were discovered to bring about much larger changes in resistance, called giant magnetoresistance. Information can then be recorded magnetically based on the direction in which the iron layer is magnetized. As a result of the discovery of giant magnetoresistance and its applications to digital storage, the 2007 Nobel Prize in Physics was awarded to Albert Fert from France and Peter Grunberg from Germany.



Engineering technology like computer storage would not be possible without a deep understanding of magnetism. (credit: Klaus Eifert)

All electric motors—with uses as diverse as powering refrigerators, starting cars, and moving elevators—contain magnets. Generators, whether producing hydroelectric power or running bicycle lights, use magnetic fields. Recycling facilities employ magnets to separate iron from other refuse. Research into using magnetic containment of fusion as a future energy source has been continuing for several years. Magnetic resonance imaging (MRI) has become an important diagnostic tool in the field of medicine, and the use of magnetism to explore brain activity is a subject of contemporary research and development. The list of applications also includes computer hard drives, tape recording, detection of inhaled asbestos, and levitation of high-speed trains. Magnetism is involved in the structure of atomic energy levels, as well as the motion of cosmic rays and charged particles trapped in the Van Allen belts around Earth. Once again, we see that all these disparate phenomena are linked by a small number of underlying physical principles.

Summary

- Magnets have two types of magnetic poles, called the north magnetic pole and the south magnetic pole. North magnetic poles are those that are attracted toward Earth’s geographic North Pole.
- Like poles repel and unlike poles attract.
- Discoveries of how magnets respond to currents by Oersted and others created a framework that led to the invention of modern electronic devices, electric motors, and magnetic imaging technology.

Glossary

north magnetic pole
: currently where a compass points to north, near the geographic North Pole; this is the effective south pole of a bar magnet but has flipped between the effective north and south poles of a bar magnet multiple times over the age of Earth

south magnetic pole
: currently where a compass points to the south, near the geographic South Pole; this is the effective north pole of a bar magnet but has

flipped just like the north magnetic pole

Magnetic Fields and Lines
By the end of this section, you will be able to:

- Define the magnetic field based on a moving charge experiencing a force
- Apply the right-hand rule to determine the direction of a magnetic force based on the motion of a charge in a magnetic field
- Sketch magnetic field lines to understand which way the magnetic field points and how strong it is in a region of space

We have outlined the properties of magnets, described how they behave, and listed some of the applications of magnetic properties. Even though there are no such things as isolated magnetic charges, we can still define the attraction and repulsion of magnets as based on a field. In this section, we define the magnetic field, determine its direction based on the right-hand rule, and discuss how to draw magnetic field lines.

Defining the Magnetic Field

A magnetic field is defined by the force that a charged particle experiences moving in this field, after we account for the gravitational and any additional electric forces possible on the charge. The magnitude of this force is proportional to the amount of charge *q*, the speed of the charged particle *v*, and the magnitude of the applied magnetic field. The direction of this force is perpendicular to both the direction of the moving charged particle and the direction of the applied magnetic field. Based on these observations, we define the magnetic field strength *B* based on the **magnetic force** $\vec{\mathbf{F}}$ on a charge *q* moving at velocity $\vec{\mathbf{v}}$ as the cross product of the velocity and magnetic field, that is,

Note:
Equation:

$$\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}.$$

In fact, this is how we define the magnetic field $\vec{\mathbf{B}}$—in terms of the force on a charged particle moving in a magnetic field. The magnitude of the force is determined from the definition of the cross product as it relates to the magnitudes of each of the vectors. In other words, the magnitude of the force satisfies

Note:
Equation:

$$F = qvB\sin\theta$$

where θ is the angle between the velocity and the magnetic field.

The SI unit for magnetic field strength B is called the **tesla** (T) after the eccentric but brilliant inventor Nikola Tesla (1856–1943), where

Equation:

$$1\ \text{T} = \frac{1\ \text{N}}{\text{A} \cdot \text{m}}.$$

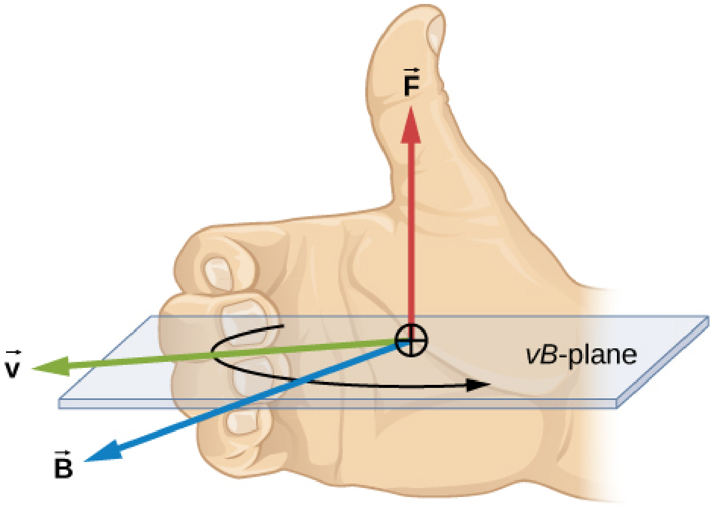
A smaller unit, called the **gauss** (G), where $1\ \text{G} = 10^{-4}\text{T}$, is sometimes used. The strongest permanent magnets have fields near 2 T; superconducting electromagnets may attain 10 T or more. Earth's magnetic field on its surface is only about $5\ \times\ 10^{-5}\text{T}$, or 0.5 G.

Note:

Problem-Solving Strategy: Direction of the Magnetic Field by the Right-Hand Rule

The direction of the magnetic force $\vec{\mathbf{F}}$ is perpendicular to the plane formed by $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$, as determined by the **right-hand rule-1** (or RHR-1), which is illustrated in [link].

1. Orient your right hand so that your fingers curl in the plane defined by the velocity and magnetic field vectors.
2. Using your right hand, sweep from the velocity toward the magnetic field with your fingers through the smallest angle possible.
3. The magnetic force is directed where your thumb is pointing.
4. If the charge was negative, reverse the direction found by these steps.



Magnetic fields exert forces on moving charges. The direction of the magnetic force on a moving charge is perpendicular to the plane formed by $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$ and

follows the right-hand rule-1 (RHR-1) as shown.
The magnitude of the force is proportional to
q, v, B, and the sine of the angle between $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$.

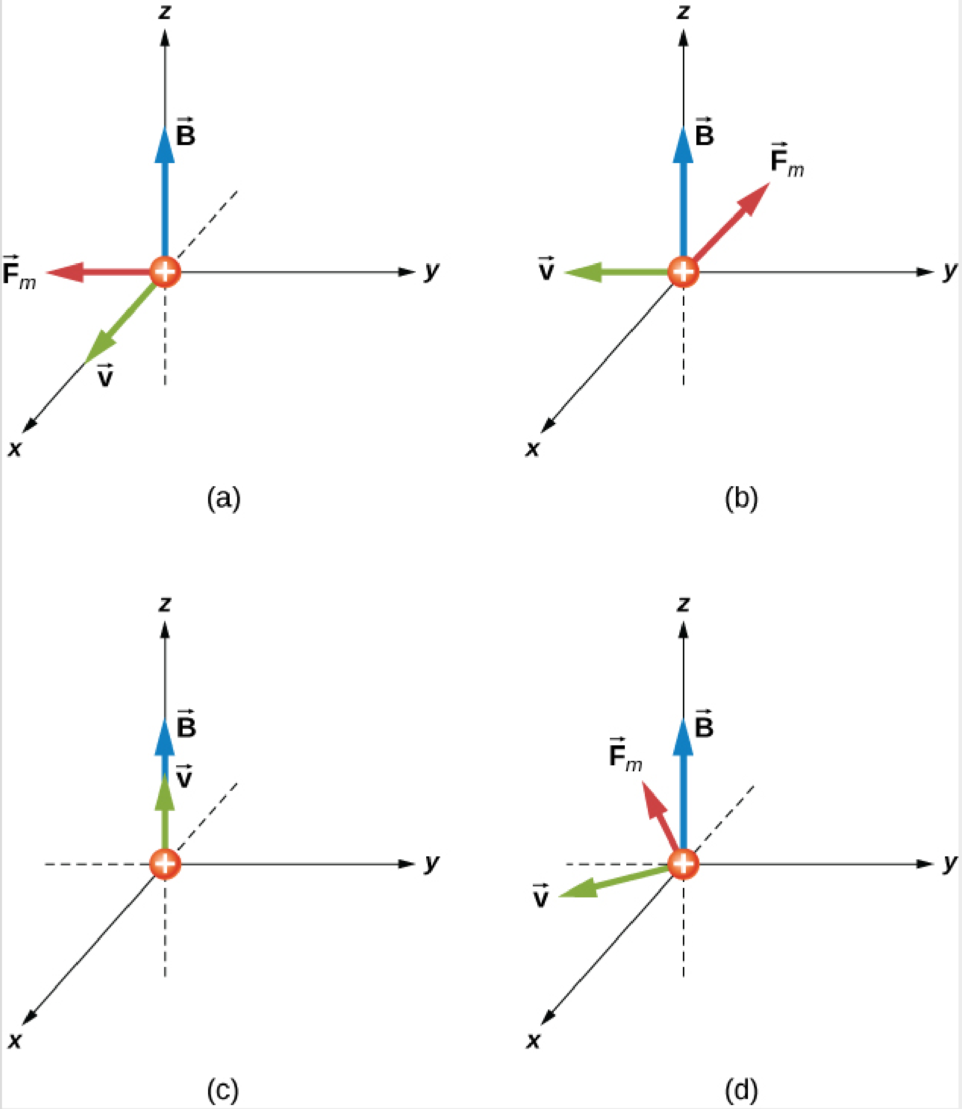
Note:
Visit this website for additional practice with the direction of magnetic fields.

There is no magnetic force on static charges. However, there is a magnetic force on charges moving at an angle to a magnetic field. When charges are stationary, their electric fields do not affect magnets. However, when charges move, they produce magnetic fields that exert forces on other magnets. When there is relative motion, a connection between electric and magnetic forces emerges—each affects the other.

Example:

An Alpha-Particle Moving in a Magnetic Field

An alpha-particle $\left(q = 3.2\ \times\ 10^{-19}\text{C}\right)$ moves through a uniform magnetic field whose magnitude is 1.5 T. The field is directly parallel to the positive *z*-axis of the rectangular coordinate system of [link]. What is the magnetic force on the alpha-particle when it is moving (a) in the positive *x*-direction with a speed of $5.0\ \times\ 10^4 \text{m/s}$? (b) in the negative *y*-direction with a speed of $5.0\ \times\ 10^4 \text{m/s}$? (c) in the positive *z*-direction with a speed of $5.0\ \times\ 10^4 \text{m/s}$? (d) with a velocity $\vec{\mathbf{v}} = \left(2.0\hat{\mathbf{i}} - 3.0\hat{\mathbf{j}} + 1.0\widehat{\mathbf{k}}\right)\ \times\ 10^4 \text{m/s}$?



The magnetic forces on an alpha-particle moving in a uniform magnetic field. The field is the same in each drawing, but the velocity is different.

Strategy
We are given the charge, its velocity, and the magnetic field strength and direction. We can thus use the equation $\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ or $F = qvB\sin\theta$ to calculate the force. The direction of the force is determined by RHR-1.

Solution

a. First, to determine the direction, start with your fingers pointing in the positive *x*-direction. Sweep your fingers upward in the direction of magnetic field. Your thumb should point in the negative *y*-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:
Equation:

$$\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}} = \left(3.2 \times 10^{-19}\text{C}\right)\left(5.0 \times 10^{4}\text{m/s}\,\hat{\mathbf{i}}\right) \times \left(1.5\,\text{T}\,\widehat{\mathbf{k}}\right) = -2.4 \times 10^{-14}\text{N}\,\hat{\mathbf{j}}.$$

b. First, to determine the directionality, start with your fingers pointing in the negative *y*-direction. Sweep your fingers upward in the direction of magnetic field as in the previous problem. Your thumb should be open in the negative *x*-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:
Equation:

$$\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}} = \left(3.2 \times 10^{-19}\text{C}\right)\left(-5.0 \times 10^{4}\text{m/s}\,\hat{\mathbf{j}}\right) \times \left(1.5\,\text{T}\,\widehat{\mathbf{k}}\right) = -2.4 \times 10^{-14}\text{N}\,\hat{\mathbf{i}}.$$

An alternative approach is to use [link] to find the magnitude of the force. This applies for both parts (a) and (b). Since the velocity is perpendicular to the magnetic field, the angle between them is 90 degrees. Therefore, the magnitude of the force is:
Equation:

$$F = qvB\sin\theta = \left(3.2 \times 10^{-19}\text{C}\right)\left(5.0 \times 10^{4}\text{m/s}\right)\left(1.5\,\text{T}\right)\sin\left(90^{\circ}\right) = 2.4 \times 10^{-14}\text{N}.$$

c. Since the velocity and magnetic field are parallel to each other, there is no orientation of your hand that will result in a force direction. Therefore, the force on this moving charge is zero. This is confirmed by the cross product. When you cross two vectors pointing in the same direction, the result is equal to zero.

d. First, to determine the direction, your fingers could point in any orientation; however, you must sweep your fingers upward in the direction of the magnetic field. As you rotate your hand, notice that the thumb can point in any *x*- or *y*-direction possible, but not in the *z*-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:
Equation:

$$\begin{aligned}\vec{\mathbf{F}} &= q\vec{\mathbf{v}} \times \vec{\mathbf{B}} = \left(3.2 \times 10^{-19}\text{C}\right)\left(\left(2.0\hat{\mathbf{i}} - 3.0\hat{\mathbf{j}} + 1.0\widehat{\mathbf{k}}\right) \times 10^{4}\text{m/s}\right) \times \left(1.5\,\text{T}\,\widehat{\mathbf{k}}\right)\\ &= \left(-14.4\hat{\mathbf{i}} - 9.6\hat{\mathbf{j}}\right) \times 10^{-15}\text{N}.\end{aligned}$$

This solution can be rewritten in terms of a magnitude and angle in the *xy*-plane:
Equation:

$$\begin{aligned}\left|\vec{\mathbf{F}}\right| &= \sqrt{F_x^2 + F_y^2} = \sqrt{(-14.4)^2 + (-9.6)^2} \times 10^{-15}\text{N} = 1.7 \times 10^{-14}\text{N}\\ \theta &= \tan^{-1}\left(\frac{F_y}{F_x}\right) = \tan^{-1}\left(\frac{-9.6 \times 10^{-15}\text{N}}{-14.4 \times 10^{-15}\text{N}}\right) = 34^{\circ}.\end{aligned}$$

The magnitude of the force can also be calculated using [link]. The velocity in this question, however, has three components. The *z*-component of the velocity can be neglected, because it is parallel to the magnetic field and therefore generates no force. The magnitude of the velocity is calculated from the *x*- and *y*-components. The angle between the velocity in the *xy*-plane and the magnetic field in the *z*-plane is 90 degrees. Therefore, the force is calculated to be:

Equation:

$$\begin{aligned} |\vec{\mathbf{v}}| &= \sqrt{(2)^2 + (-3)^2} \times 10^4 \frac{\text{m}}{\text{s}} = 3.6 \times 10^4 \frac{\text{m}}{\text{s}} \\ F &= qvB\sin\theta = (3.2 \times 10^{-19}\text{C})(3.6 \times 10^4 \text{m/s})(1.5\text{ T})\sin(90^\circ) = 1.7 \times 10^{-14}\text{N}. \end{aligned}$$

This is the same magnitude of force calculated by unit vectors.

Significance

The cross product in this formula results in a third vector that must be perpendicular to the other two. Other physical quantities, such as angular momentum, also have three vectors that are related by the cross product. Note that typical force values in magnetic force problems are much larger than the gravitational force. Therefore, for an isolated charge, the magnetic force is the dominant force governing the charge's motion.

Note:

Exercise:

Problem:

Check Your Understanding Repeat the previous problem with the magnetic field in the *x*-direction rather than in the *z*-direction. Check your answers with RHR-1.

Solution:

a. 0 N; b. $2.4 \times 10^{-14}\hat{\mathbf{k}}$N; c. $2.4 \times 10^{-14}\hat{\mathbf{j}}$ N; d. $(7.2\hat{\mathbf{j}} + 2.2\hat{\mathbf{k}}) \times 10^{-15}$N

Representing Magnetic Fields

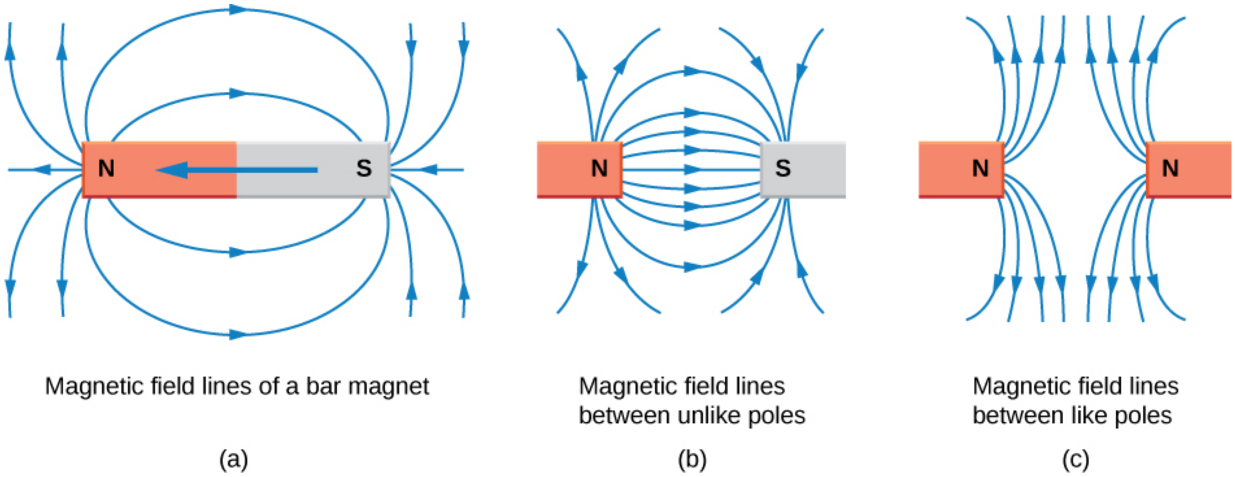
The representation of magnetic fields by **magnetic field lines** is very useful in visualizing the strength and direction of the magnetic field. As shown in [link], each of these lines forms a closed loop, even if not shown by the constraints of the space available for the figure. The field lines emerge from the north pole (N), loop around to the south pole (S), and continue through the bar magnet back to the north pole.

Magnetic field lines have several hard-and-fast rules:

1. The direction of the magnetic field is tangent to the field line at any point in space. A small compass will point in the direction of the field line.
2. The strength of the field is proportional to the closeness of the lines. It is exactly proportional to the number of lines per unit area perpendicular to the lines (called the areal density).
3. Magnetic field lines can never cross, meaning that the field is unique at any point in space.

4. Magnetic field lines are continuous, forming closed loops without a beginning or end. They are directed from the north pole to the south pole.

The last property is related to the fact that the north and south poles cannot be separated. It is a distinct difference from electric field lines, which generally begin on positive charges and end on negative charges or at infinity. If isolated magnetic charges (referred to as magnetic monopoles) existed, then magnetic field lines would begin and end on them.



Magnetic field lines of a bar magnet

(a)

Magnetic field lines between unlike poles

(b)

Magnetic field lines between like poles

(c)

Magnetic field lines are defined to have the direction in which a small compass points when placed at a location in the field. The strength of the field is proportional to the closeness (or density) of the lines. If the interior of the magnet could be probed, the field lines would be found to form continuous, closed loops. To fit in a reasonable space, some of these drawings may not show the closing of the loops; however, if enough space were provided, the loops would be closed.

Summary

- Charges moving across a magnetic field experience a force determined by $\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$. The force is perpendicular to the plane formed by $\vec{\mathbf{v}}$ and $\vec{\mathbf{B}}$.
- The direction of the force on a moving charge is given by the right hand rule 1 (RHR-1): Sweep your fingers in a velocity, magnetic field plane. Start by pointing them in the direction of velocity and sweep towards the magnetic field. Your thumb points in the direction of the magnetic force for positive charges.
- Magnetic fields can be pictorially represented by magnetic field lines, which have the following properties:
    1. The field is tangent to the magnetic field line.
    2. Field strength is proportional to the line density.
    3. Field lines cannot cross.
    4. Field lines form continuous, closed loops.

- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

Conceptual Questions

Exercise:

Problem:

Discuss the similarities and differences between the electrical force on a charge and the magnetic force on a charge.

Solution:

Both are field dependent. Electrical force is dependent on charge, whereas magnetic force is dependent on current or rate of charge flow.

Exercise:

Problem:

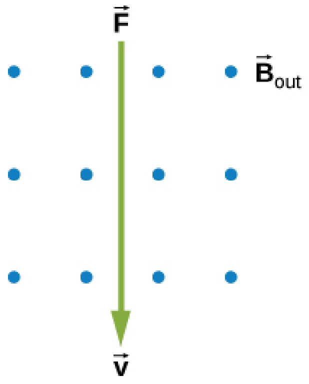
(a) Is it possible for the magnetic force on a charge moving in a magnetic field to be zero? (b) Is it possible for the electric force on a charge moving in an electric field to be zero? (c) Is it possible for the resultant of the electric and magnetic forces on a charge moving simultaneously through both fields to be zero?

Problems

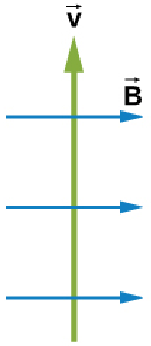
Exercise:

Problem:

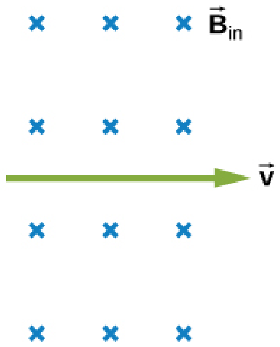
What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases?


$\vec{F}$
$\vec{B}_{out}$
$\vec{v}$

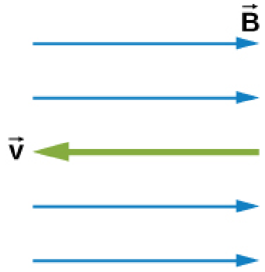
(a)



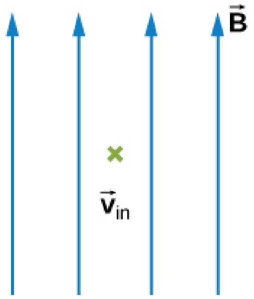
(b)


$\vec{B}_{in}$
$\vec{v}$

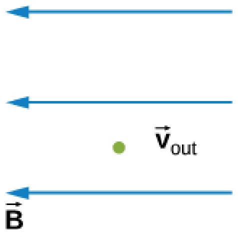
(c)



(d)


$\vec{B}$
$\vec{v}_{in}$

(e)


$\vec{v}_{out}$
$\vec{B}$

(f)

Solution:

a. left; b. into the page; c. up the page; d. no force; e. right; f. down

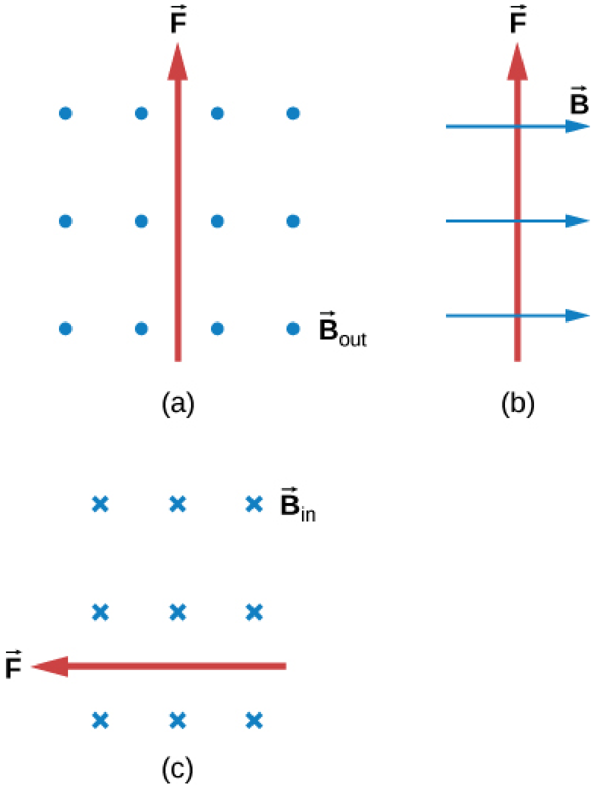
Exercise:

Problem: Repeat previous exercise for a negative charge.

Exercise:

Problem:

What is the direction of the velocity of a negative charge that experiences the magnetic force shown in each of the three cases, assuming it moves perpendicular to *B*?



Solution:

a. right; b. into the page; c. down

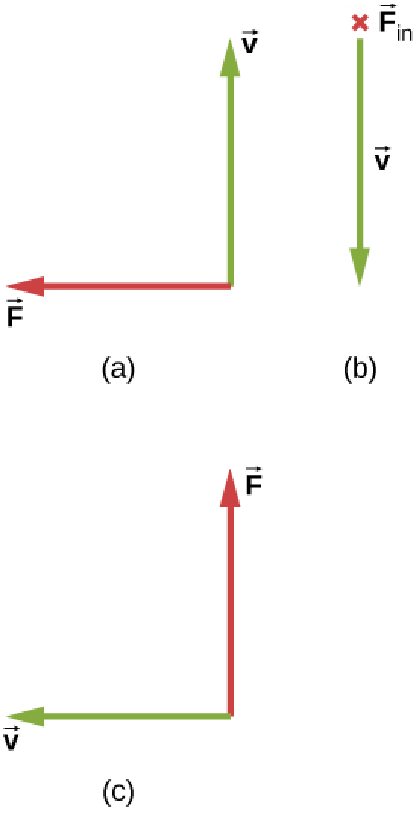
Exercise:

Problem: Repeat previous exercise for a positive charge.

Exercise:

Problem:

What is the direction of the magnetic field that produces the magnetic force on a positive charge as shown in each of the three cases, assuming $\vec{B}$ is perpendicular to $\vec{v}$?



(a) (b) (c)

Solution:

a. into the page; b. left; c. out of the page

Exercise:

Problem: Repeat previous exercise for a negative charge.

Exercise:

Problem:

(a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a 0.500-μC charge and flies due west at a speed of 660. m/s over Earth's south magnetic pole, where the $8.00 \times 10^{-5} - \text{T}$ magnetic field points straight down into the ground. What are the direction and the magnitude of the magnetic force on the plane? (b) Discuss whether the value obtained in part (a) implies this is a significant or negligible effect.

Solution:

a. 2.64×10^{-8} N; north b. The force is very small, so this implies that the effect of static charges on airplanes is negligible.

Exercise:

Problem:

(a) A cosmic ray proton moving toward Earth at $5.00 \times 10^{7} \mathrm{m/s}$ experiences a magnetic force of $1.70 \times 10^{-16}\,\mathrm{N}$. What is the strength of the magnetic field if there is a 45º angle between it and the proton's velocity? (b) Is the value obtained in part a. consistent with the known strength of Earth's magnetic field on its surface? Discuss.

Exercise:

Problem:

An electron moving at $4.00 \times 10^{3} \mathrm{m/s}$ in a 1.25-T magnetic field experiences a magnetic force of $1.40 \times 10^{-16}\,\mathrm{N}$. What angle does the velocity of the electron make with the magnetic field? There are two answers.

Solution:

$10.1^{\circ}; 169.9^{\circ}$

Exercise:

Problem:

(a) A physicist performing a sensitive measurement wants to limit the magnetic force on a moving charge in her equipment to less than $1.00 \times 10^{-12}\,\mathrm{N}$. What is the greatest the charge can be if it moves at a maximum speed of 30.0 m/s in Earth's field? (b) Discuss whether it would be difficult to limit the charge to less than the value found in (a) by comparing it with typical static electricity and noting that static is often absent.

Glossary

gauss
: G, unit of the magnetic field strength; $1\ \mathrm{G} = 10^{-4}\mathrm{T}$

magnetic field lines
: continuous curves that show the direction of a magnetic field; these lines point in the same direction as a compass points, toward the magnetic south pole of a bar magnet

magnetic force
: force applied to a charged particle moving through a magnetic field

right-hand rule-1
: using your right hand to determine the direction of either the magnetic force, velocity of a charged particle, or magnetic field

tesla
: SI unit for magnetic field: 1 T = 1 N/A-m

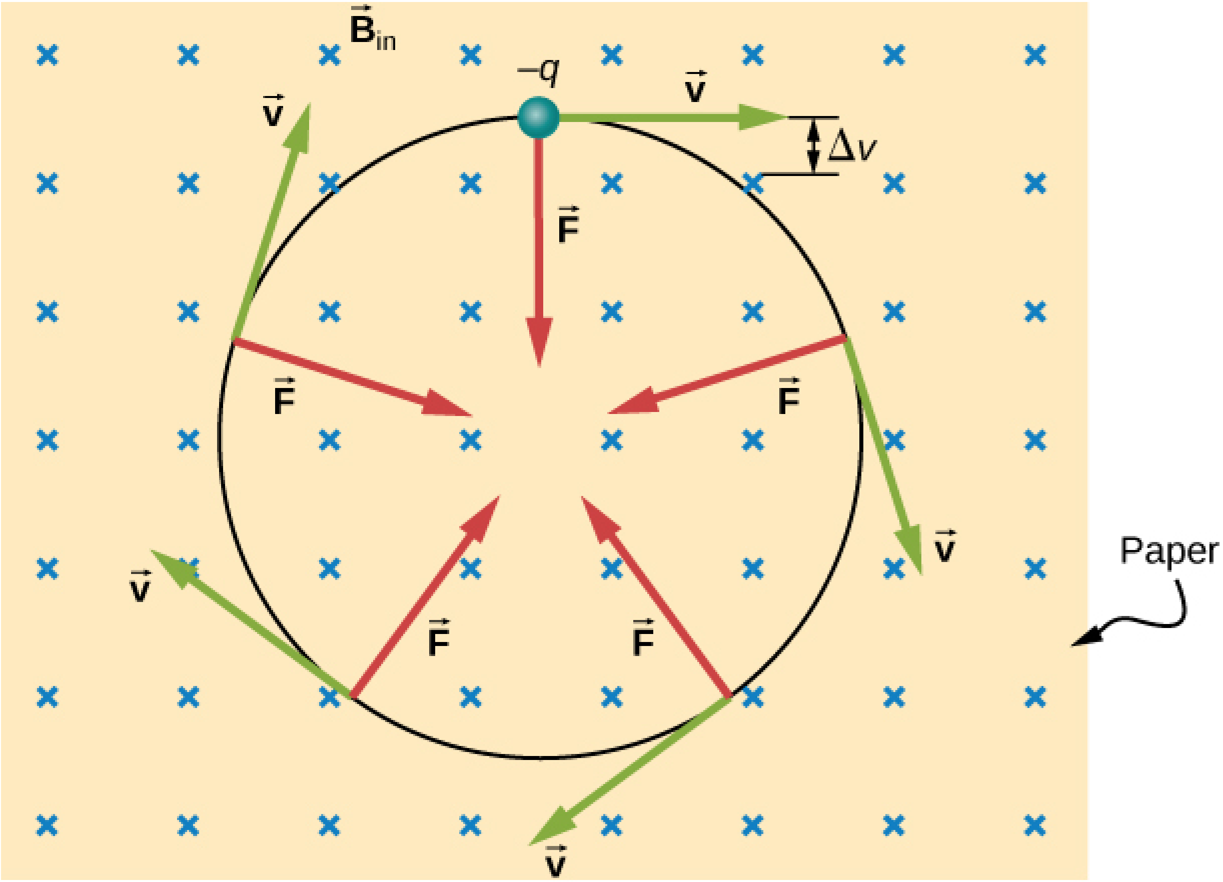
Motion of a Charged Particle in a Magnetic Field

By the end of this section, you will be able to:

- Explain how a charged particle in an external magnetic field undergoes circular motion
- Describe how to determine the radius of the circular motion of a charged particle in a magnetic field

A charged particle experiences a force when moving through a magnetic field. What happens if this field is uniform over the motion of the charged particle? What path does the particle follow? In this section, we discuss the circular motion of the charged particle as well as other motion that results from a charged particle entering a magnetic field.

The simplest case occurs when a charged particle moves perpendicular to a uniform *B*-field ([link]). If the field is in a vacuum, the magnetic field is the dominant factor determining the motion. Since the magnetic force is perpendicular to the direction of travel, a charged particle follows a curved path in a magnetic field. The particle continues to follow this curved path until it forms a complete circle. Another way to look at this is that the magnetic force is always perpendicular to velocity, so that it does no work on the charged particle. The particle’s kinetic energy and speed thus remain constant. The direction of motion is affected but not the speed.



A negatively charged particle moves in the plane of the paper in a region where the magnetic field is perpendicular to the paper (represented by the small × ’s—like the tails of arrows). The magnetic force is perpendicular to the velocity, so velocity changes in direction but not magnitude. The result is uniform circular motion. (Note that because the charge is negative, the force is opposite in direction to the prediction of the right-hand rule.)

In this situation, the magnetic force supplies the centripetal force $F_{\text{c}} = \frac{mv^2}{r}$. Noting that the velocity is perpendicular to the magnetic field, the magnitude of the magnetic force is reduced to $F = qvB$. Because the magnetic force F supplies the centripetal force F_c, we have

Equation:

$$qvB = \frac{mv^2}{r}.$$

Solving for *r* yields

Note:
Equation:

$$r = \frac{mv}{qB}.$$

Here, *r* is the radius of curvature of the path of a charged particle with mass *m* and charge *q*, moving at a speed *v* that is perpendicular to a magnetic field of strength *B*. The time for the charged particle to go around the circular path is defined as the period, which is the same as the distance traveled (the circumference) divided by the speed. Based on this and [link], we can derive the period of motion as

Note:
Equation:

$$T = \frac{2\pi r}{v} = \frac{2\pi}{v}\frac{mv}{qB} = \frac{2\pi m}{qB}.$$

If the velocity is not perpendicular to the magnetic field, then we can compare each component of the velocity separately with the magnetic field. The component of the velocity perpendicular to the magnetic field produces a magnetic force perpendicular to both this velocity and the field:

Equation:

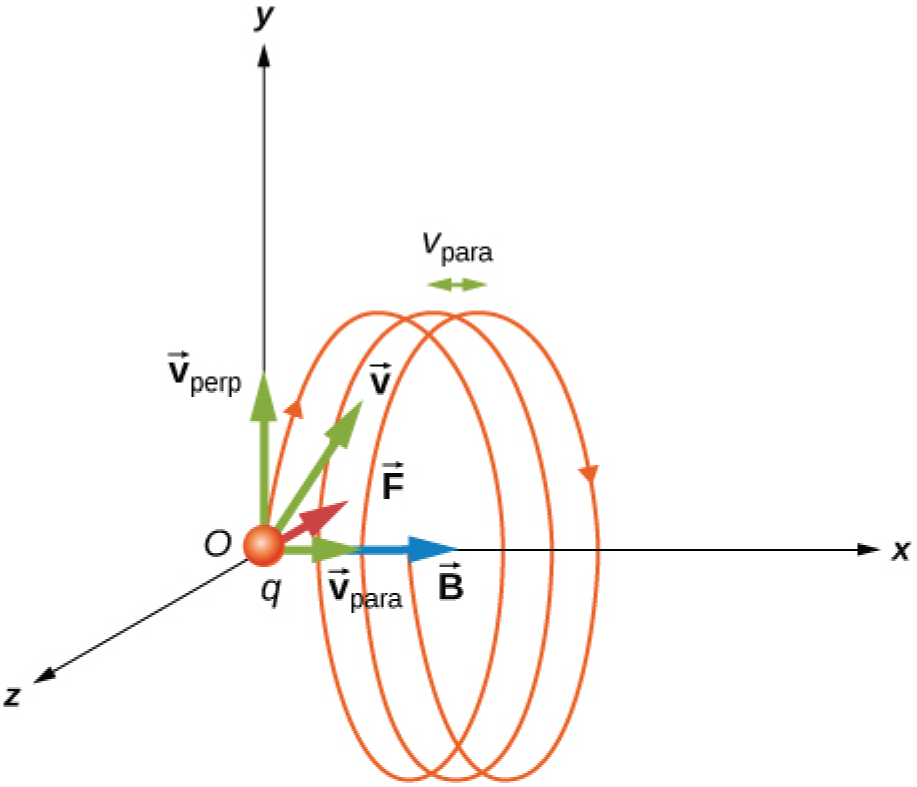
$$v_{\text{perp}} = v\sin\theta, \quad v_{\text{para}} = v\cos\theta.$$

where θ is the angle between *v* and *B*. The component parallel to the magnetic field creates constant motion along the same direction as the magnetic field, also shown in [link]. The parallel motion determines the *pitch p* of the helix, which is the distance between adjacent turns. This distance equals the parallel component of the velocity times the period:

Equation:

$$p = v_{\text{para}}T.$$

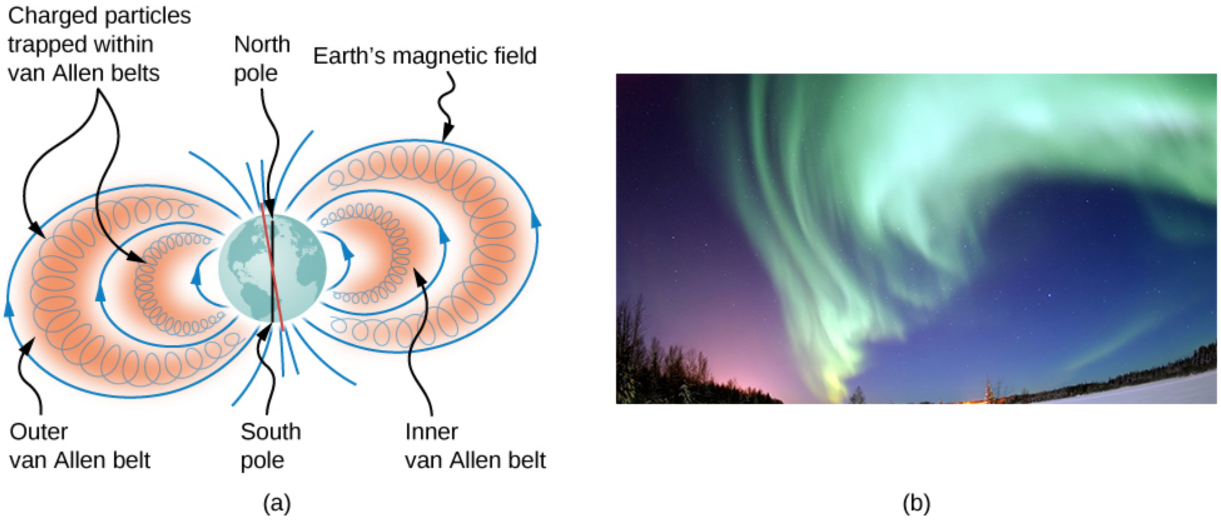
The result is a **helical motion**, as shown in the following figure.



A charged particle moving with a velocity not in the same direction as the magnetic field. The velocity component perpendicular to the magnetic field creates circular motion, whereas the component of the velocity parallel to the field moves the particle along a straight line. The pitch is the horizontal distance between two consecutive circles. The resulting motion is helical.

While the charged particle travels in a helical path, it may enter a region where the magnetic field is not uniform. In particular, suppose a particle travels from a region of strong magnetic field to a region of weaker field, then back to a region of stronger field. The particle may reflect back before entering the stronger magnetic field region. This is similar to a wave on a string traveling from a very light, thin string to a hard wall and reflecting backward. If the reflection happens at both ends, the particle is trapped in a so-called magnetic bottle.

Trapped particles in magnetic fields are found in the Van Allen radiation belts around Earth, which are part of Earth’s magnetic field. These belts were discovered by James Van Allen while trying to measure the flux of **cosmic rays** on Earth (high-energy particles that come from outside the solar system) to see whether this was similar to the flux measured on Earth. Van Allen found that due to the contribution of particles trapped in Earth’s magnetic field, the flux was much higher on Earth than in outer space. Aurorae, like the famous aurora borealis (northern lights) in the Northern Hemisphere ([link]), are beautiful displays of light emitted as ions recombine with electrons entering the atmosphere as they spiral along magnetic field lines. (The ions are primarily oxygen and nitrogen atoms that are initially ionized by collisions with energetic particles in Earth’s atmosphere.) Aurorae have also been observed on other planets, such as Jupiter and Saturn.

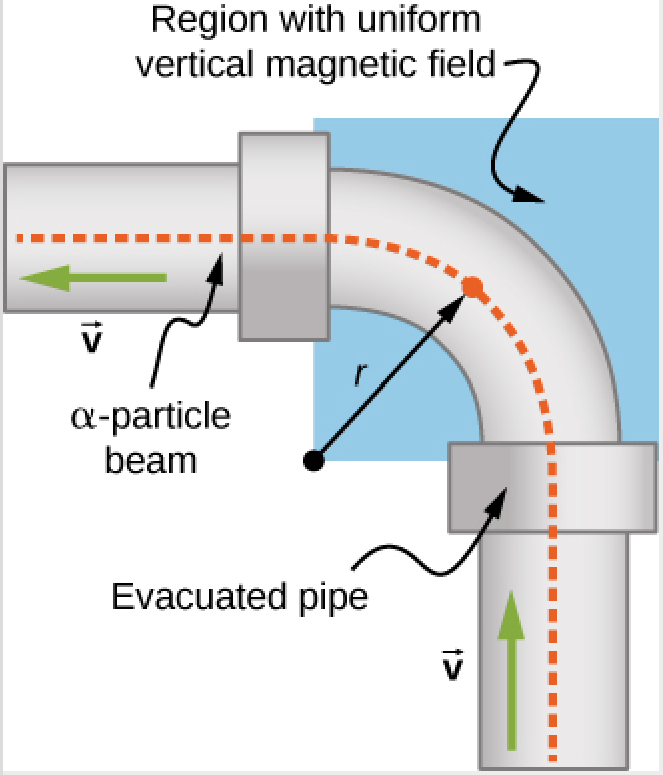

Charged particles trapped within van Allen belts
North pole
Earth's magnetic field
Outer van Allen belt
South pole
Inner van Allen belt
(a)
(b)

(a) The Van Allen radiation belts around Earth trap ions produced by cosmic rays striking Earth’s atmosphere. (b) The magnificent spectacle of the aurora borealis, or northern lights, glows in the northern sky above Bear Lake near Eielson Air Force Base, Alaska. Shaped by Earth’s magnetic field, this light is produced by glowing molecules and ions of oxygen and nitrogen. (credit b: modification of work by USAF Senior Airman Joshua Strang)

Example:

Beam Deflector

A research group is investigating short-lived radioactive isotopes. They need to design a way to transport alpha-particles (helium nuclei) from where they are made to a place where they will collide with another material to form an isotope. The beam of alpha-particles $\left(m=6.64\times 10^{-27}\text{kg},q=3.2\times 10^{-19}\text{C}\right)$ bends through a 90-degree region with a uniform magnetic field of 0.050 T ([link]). (a) In what direction should the magnetic field be applied? (b) How much time does it take the alpha-particles to traverse the uniform magnetic field region?


Region with uniform vertical magnetic field
$\vec{v}$
α-particle beam
r
Evacuated pipe
$\vec{v}$

Top view of the beam deflector setup.

Strategy

a. The direction of the magnetic field is shown by the RHR-1. Your fingers point in the direction of *v*, and your thumb needs to point in the direction of the force, to the left. Therefore, since the alpha-particles are positively charged, the magnetic field must point down.
b. The period of the alpha-particle going around the circle is

Note:
Equation:

$$T = \frac{2\pi m}{qB}.$$

Because the particle is only going around a quarter of a circle, we can take 0.25 times the period to find the time it takes to go around this path.

Solution

a. Let’s start by focusing on the alpha-particle entering the field near the bottom of the picture. First, point your thumb up the page. In order for your palm to open to the left where the centripetal force (and hence the magnetic force) points, your fingers need to change orientation until they point into the page. This is the direction of the applied magnetic field.
b. The period of the charged particle going around a circle is calculated by using the given mass, charge, and magnetic field in the problem. This works out to be
Equation:

$$T = \frac{2\pi m}{qB} = \frac{2\pi \left(6.64 \times 10^{-27}\text{kg}\right)}{\left(3.2 \times 10^{-19}\text{C}\right)(0.050\ \text{T})} = 2.6 \times 10^{-6}\text{s}.$$

However, for the given problem, the alpha-particle goes around a quarter of the circle, so the time it takes would be
Equation:

$$t = 0.25 \times 2.61 \times 10^{-6}\text{s} = 6.5 \times 10^{-7}\text{s}.$$

Significance
This time may be quick enough to get to the material we would like to bombard, depending on how short-lived the radioactive isotope is and continues to emit alpha-particles. If we could increase the magnetic field applied in the region, this would shorten the time even more. The path the

particles need to take could be shortened, but this may not be economical given the experimental setup.

Note:
Exercise:

Problem:

Check Your Understanding A uniform magnetic field of magnitude 1.5 T is directed horizontally from west to east. (a) What is the magnetic force on a proton at the instant when it is moving vertically downward in the field with a speed of $4 \times 10^7 \text{ m/s}$? (b) Compare this force with the weight *w* of a proton.

Solution:

a. 9.6×10^{-12}N toward the south; b. $\frac{w}{F_\text{m}} = 1.7 \times 10^{-15}$

Example:
Helical Motion in a Magnetic Field

A proton enters a uniform magnetic field of 1.0×10^{-4}T with a speed of $5 \times 10^5 \text{ m/s}$. At what angle must the magnetic field be from the velocity so that the pitch of the resulting helical motion is equal to the radius of the helix?

Strategy

The pitch of the motion relates to the parallel velocity times the period of the circular motion, whereas the radius relates to the perpendicular velocity component. After setting the radius and the pitch equal to each other, solve for the angle between the magnetic field and velocity or θ.

Solution

The pitch is given by [link], the period is given by [link], and the radius of circular motion is given by [link]. Note that the velocity in the radius equation is related to only the perpendicular velocity, which is where the circular motion occurs. Therefore, we substitute the sine component of the overall velocity into the radius equation to equate the pitch and radius:

Equation:

$$\begin{aligned} p &= r \\ v_{\|}T &= \frac{mv_{\perp}}{qB} \\ v\cos\theta\frac{2\pi m}{qB} &= \frac{mv\sin\theta}{qB} \\ 2\pi &= \tan\theta \\ \theta &= 81.0^{\circ}. \end{aligned}$$

Significance

If this angle were 0°, only parallel velocity would occur and the helix would not form, because there would be no circular motion in the perpendicular plane. If this angle were 90°, only circular motion would occur and there would be no movement of the circles perpendicular to the motion. That is what creates the helical motion.

Summary

- A magnetic force can supply centripetal force and cause a charged particle to move in a circular path of radius $r = \frac{mv}{qB}$.
- The period of circular motion for a charged particle moving in a magnetic field perpendicular to the plane of motion is $T = \frac{2\pi m}{qB}$.
- Helical motion results if the velocity of the charged particle has a component parallel to the magnetic field as well as a component perpendicular to the magnetic field.

Conceptual Questions

Exercise:

Problem:

At a given instant, an electron and a proton are moving with the same velocity in a constant magnetic field. Compare the magnetic forces on these particles. Compare their accelerations.

Solution:

The magnitude of the proton and electron magnetic forces are the same since they have the same amount of charge. The direction of these forces however are opposite of each other. The accelerations are opposite in direction and the electron has a larger acceleration than the proton due to its smaller mass.

Exercise:

Problem:

Does increasing the magnitude of a uniform magnetic field through which a charge is traveling necessarily mean increasing the magnetic force on the charge? Does changing the direction of the field necessarily mean a change in the force on the charge?

Exercise:

Problem:

An electron passes through a magnetic field without being deflected. What do you conclude about the magnetic field?

Solution:

The magnetic field must point parallel or anti-parallel to the velocity.

Exercise:

Problem:

If a charged particle moves in a straight line, can you conclude that there is no magnetic field present?

Exercise:

Problem:

How could you determine which pole of an electromagnet is north and which pole is south?

Solution:

A compass points toward the north pole of an electromagnet.

Problems

Exercise:

Problem:

A cosmic-ray electron moves at $7.5 \times 10^6 \mathrm{m/s}$ perpendicular to Earth’s magnetic field at an altitude where the field strength is $1.0 \times 10^{-5} \mathrm{T}$. What is the radius of the circular path the electron follows?

Solution:

4.27 m

Exercise:

Problem:

(a) Viewers of Star Trek have heard of an antimatter drive on the Starship *Enterprise*. One possibility for such a futuristic energy source is to store antimatter charged particles in a vacuum chamber, circulating in a magnetic field, and then extract them as needed. Antimatter annihilates normal matter, producing pure energy. What strength magnetic field is needed to hold antiprotons, moving at $5.0 \times 10^7 \mathrm{m/s}$ in a circular path 2.00 m in radius? Antiprotons have the same mass as protons but the opposite (negative) charge. (b) Is this field strength obtainable with today's technology or is it a futuristic possibility?

Exercise:

Problem:

(a) An oxygen-16 ion with a mass of $2.66 \times 10^{-26} \mathrm{kg}$ travels at $5.0 \times 10^6 \mathrm{m/s}$ perpendicular to a 1.20-T magnetic field, which makes it move in a circular arc with a 0.231-m radius. What positive charge is on the ion? (b) What is the ratio of this charge to the charge of an electron? (c) Discuss why the ratio found in (b) should be an integer.

Solution:

a. 4.80×10^{-19} C; b. 3; c. This ratio must be an integer because charges must be integer numbers of the basic charge of an electron. There are no free charges with values less than this basic charge, and all charges are integer multiples of this basic charge.

Exercise:

Problem:

An electron in a TV CRT moves with a speed of $6.0 \times 10^{6} \mathrm{m/s}$, in a direction perpendicular to Earth’s field, which has a strength of $5.0 \times 10^{-5} \mathrm{T}$. (a) What strength electric field must be applied perpendicular to the Earth’s field to make the electron moves in a straight line? (b) If this is done between plates separated by 1.00 cm, what is the voltage applied? (Note that TVs are usually surrounded by a ferromagnetic material to shield against external magnetic fields and avoid the need for such a correction.)

Exercise:

Problem:

(a) At what speed will a proton move in a circular path of the same radius as the electron in the previous exercise? (b) What would the radius of the path be if the proton had the same speed as the electron? (c) What would the radius be if the proton had the same kinetic energy as the electron? (d) The same momentum?

Solution:

(a) 3.27 x 10^{4} m/s (b) 12,525 m (c) 292 m (d) 6.83 m.

Exercise:

Problem:

(a) What voltage will accelerate electrons to a speed of $6.00 \times 10^{-7} \mathrm{m/s}$? (b) Find the radius of curvature of the path of a proton accelerated through this potential in a 0.500-T field and compare this with the radius of curvature of an electron accelerated through the same potential.

Exercise:

Problem:

An alpha-particle $\left(m = 6.64 \times 10^{-27}\,\text{kg}, q = 3.2 \times 10^{-19}\,\text{C}\right)$ travels in a circular path of radius 25 cm in a uniform magnetic field of magnitude 1.5 T. (a) What is the speed of the particle? (b) What is the kinetic energy in electron-volts? (c) Through what potential difference must the particle be accelerated in order to give it this kinetic energy?

Solution:

a. $1.8 \times 10^7 \text{m/s}$; b. $6.8 \times 10^6 \text{eV}$; c. $3.4 \times 10^6 \text{V}$

Exercise:

Problem:

A particle of charge *q* and mass *m* is accelerated from rest through a potential difference *V*, after which it encounters a uniform magnetic field *B*. If the particle moves in a plane perpendicular to *B*, what is the radius of its circular orbit?

Glossary

cosmic rays	comprised of particles that originate mainly from outside the solar system and reach Earth
helical motion	superposition of circular motion with a straight-line motion that is followed by a charged particle moving in a region of magnetic field at an angle to the field

Magnetic Force on a Current-Carrying Conductor
By the end of this section, you will be able to:

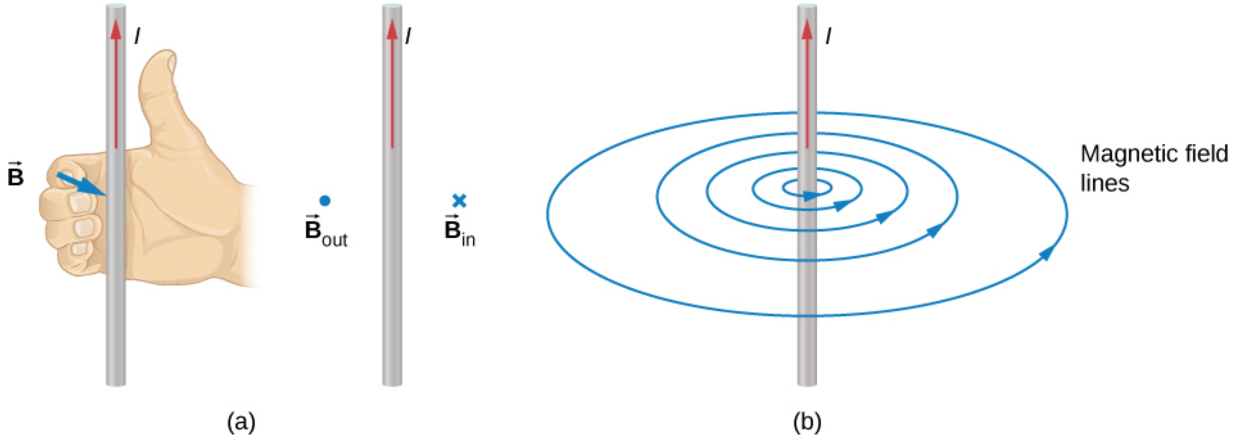
- Determine the direction in which a current-carrying wire experiences a force in an external magnetic field
- Calculate the force on a current-carrying wire in an external magnetic field

Moving charges experience a force in a magnetic field. If these moving charges are in a wire—that is, if the wire is carrying a current—the wire should also experience a force. However, before we discuss the force exerted on a current by a magnetic field, we first examine the magnetic field generated by an electric current. We are studying two separate effects here that interact closely: A current-carrying wire generates a magnetic field and the magnetic field exerts a force on the current-carrying wire.

Magnetic Fields Produced by Electrical Currents

When discussing historical discoveries in magnetism, we mentioned Oersted’s finding that a wire carrying an electrical current caused a nearby compass to deflect. A connection was established that electrical currents produce magnetic fields. (This connection between electricity and magnetism is discussed in more detail in Sources of Magnetic Fields.)

The compass needle near the wire experiences a force that aligns the needle tangent to a circle around the wire. Therefore, a current-carrying wire produces circular loops of magnetic field. To determine the direction of the magnetic field generated from a wire, we use a second right-hand rule. In RHR-2, your thumb points in the direction of the current while your fingers wrap around the wire, pointing in the direction of the magnetic field produced ([link]). If the magnetic field were coming at you or out of the page, we represent this with a dot. If the magnetic field were going into the page, we represent this with an $\times$. These symbols come from considering a vector arrow: An arrow pointed toward you, from your perspective, would look like a dot or the tip of an arrow. An arrow pointed away from you, from your perspective, would look like a cross or an $\times$. A composite sketch of the magnetic circles is shown in [link], where the field strength is shown to decrease as you get farther from the wire by loops that are farther separated.



(a) When the wire is in the plane of the paper, the field is perpendicular to the paper. Note the symbols used for the field pointing inward (like the tail of an arrow) and the field pointing outward (like the tip of an arrow). (b) A long and straight wire creates a field with magnetic field lines forming circular loops.

Calculating the Magnetic Force

Electric current is an ordered movement of charge. A current-carrying wire in a magnetic field must therefore experience a force due to the field. To investigate this force, let's consider the infinitesimal section of wire as shown in [link]. The length and cross-sectional area of the section are *dl* and *A*, respectively, so its volume is $V = A \cdot dl$. The wire is formed from material that contains *n* charge carriers per unit volume, so the number of charge carriers in the section is $nA \cdot dl$. If the charge carriers move with drift velocity $\vec{\mathbf{v}}_{\text{d}}$, the current *I* in the wire is (from Current and Resistance)

Equation:

$$I = neAv_d.$$

The magnetic force on any single charge carrier is $e\vec{\mathbf{v}}_{\text{d}} \times \vec{\mathbf{B}}$, so the total magnetic force $d\vec{\mathbf{F}}$ on the $nA \cdot dl$ charge carriers in the section of wire is

Equation:

$$d\overrightarrow{\mathbf{F}} = (nA \cdot dl)e\vec{\mathbf{v}}_{\mathrm{d}} \times \overrightarrow{\mathbf{B}}.$$

We can define *dl* to be a vector of length *dl* pointing along $\vec{\mathbf{v}}_{\mathrm{d}}$, which allows us to rewrite this equation as

Equation:

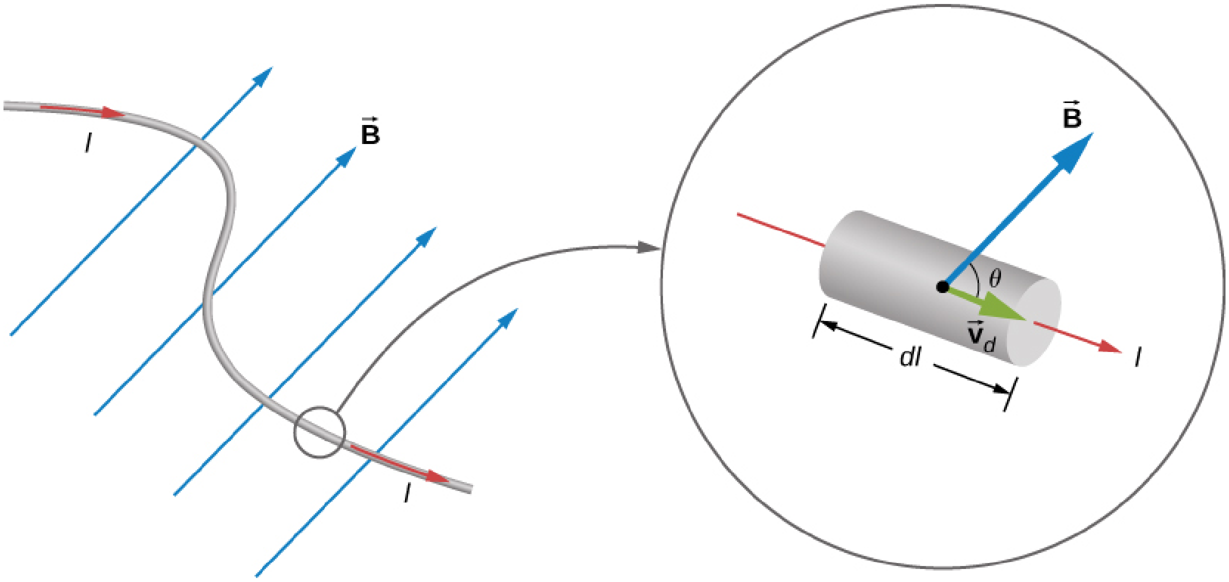
$$d\overrightarrow{\mathbf{F}} = neAv_{\mathrm{d}}\mathbf{dl} \times \overrightarrow{\mathbf{B}},$$

or

Equation:

$$d\overrightarrow{\mathbf{F}} = I\mathbf{dl} \times \overrightarrow{\mathbf{B}}.$$

This is the magnetic force on the section of wire. Note that it is actually the net force exerted by the field on the charge carriers themselves. The direction of this force is given by RHR-1, where you point your fingers in the direction of the current and curl them toward the field. Your thumb then points in the direction of the force.



An infinitesimal section of current-carrying wire in a magnetic field.

To determine the magnetic force $\vec{\mathbf{F}}$ on a wire of arbitrary length and shape, we must integrate [link] over the entire wire. If the wire section happens to be straight and B is uniform, the equation differentials become absolute quantities, giving us

Note:
Equation:

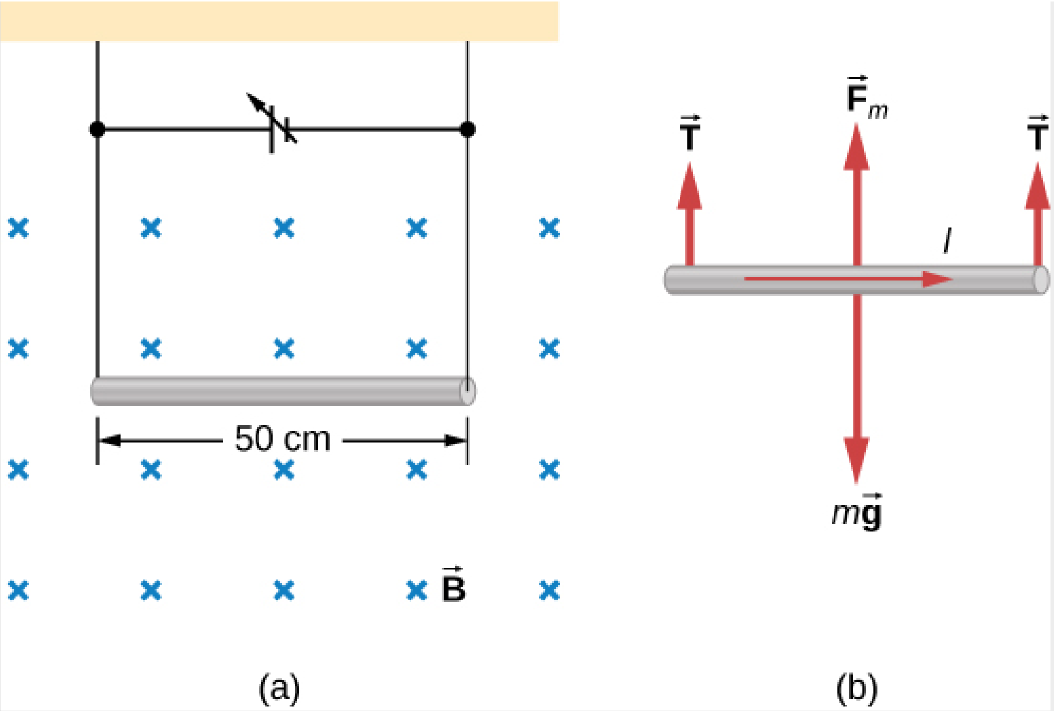
$$\vec{\mathbf{F}} = I\vec{\mathbf{l}} \times \vec{\mathbf{B}}.$$

This is the force on a straight, current-carrying wire in a uniform magnetic field.

Example:

Balancing the Gravitational and Magnetic Forces on a Current-Carrying Wire

A wire of length 50 cm and mass 10 g is suspended in a horizontal plane by a pair of flexible leads ([link]). The wire is then subjected to a constant magnetic field of magnitude 0.50 T, which is directed as shown. What are the magnitude and direction of the current in the wire needed to remove the tension in the supporting leads?



(a) A wire suspended in a magnetic field. (b) The free-body diagram for the wire.

Strategy

From the free-body diagram in the figure, the tensions in the supporting leads go to zero when the gravitational and magnetic forces balance each other. Using the RHR-1, we find that the magnetic force points up. We can then determine the current *I* by equating the two forces.

Solution

Equate the two forces of weight and magnetic force on the wire:

Equation:

$$mg = IlB.$$

Thus,

Equation:

$$I = \frac{mg}{lB} = \frac{(0.010\,\text{kg})(9.8\,\text{m/s}^2)}{(0.50\,\text{m})(0.50\,\text{T})} = 0.39\,\text{A}.$$

Significance

This large magnetic field creates a significant force on a length of wire to counteract the weight of the wire.

Example:

Calculating Magnetic Force on a Current-Carrying Wire

A long, rigid wire lying along the *y*-axis carries a 5.0-A current flowing in the positive *y*-direction. (a) If a constant magnetic field of magnitude 0.30 T is directed along the positive *x*-axis, what is the magnetic force per unit length on the wire? (b) If a constant magnetic field of 0.30 T is directed 30 degrees from the +*x*-axis towards the +*y*-axis, what is the magnetic force per unit length on the wire?

Strategy

The magnetic force on a current-carrying wire in a magnetic field is given by $\vec{\mathbf{F}} = I\vec{\mathbf{l}} \times \vec{\mathbf{B}}$. For part a, since the current and magnetic field are perpendicular in this problem, we can simplify the formula to give us the magnitude and find the direction through the RHR-1. The angle θ is 90 degrees, which means $\sin\theta = 1$. Also, the length can be divided over to the left-hand side to find the force per unit length. For part b, the current times length is written in unit vector notation, as well as the magnetic field. After the cross product is taken, the directionality is evident by the resulting unit vector.

Solution

a. We start with the general formula for the magnetic force on a wire. We are looking for the force per unit length, so we divide by the length to bring it to the left-hand side. We also set $\sin\theta = 1$. The solution therefore is
   Equation:

$$\begin{array}{rcl} F & = & IlB\sin\theta \\ \frac{F}{l} & = & (5.0\text{ A})(0.30\text{ T}) \\ \frac{F}{l} & = & 1.5\,\text{N/m}. \end{array}$$

   Directionality: Point your fingers in the positive *y*-direction and curl your fingers in the positive *x*-direction. Your thumb will point in the $-\vec{\mathbf{k}}$ direction. Therefore, with directionality, the solution is
   Equation:

$$\frac{\vec{\mathbf{F}}}{l} = -1.5\vec{\mathbf{k}}\,\text{N/m}.$$

b. The current times length and the magnetic field are written in unit vector notation. Then, we take the cross product to find the force:
Equation:

$$\begin{aligned}\vec{\mathbf{F}} &= I\vec{\mathbf{l}} \times \vec{\mathbf{B}} = (5.0A)l\hat{\mathbf{j}} \times \left(0.30T\cos(30^\circ)\hat{\mathbf{i}} + 0.30T\sin(30^\circ)\hat{\mathbf{j}}\right) \\ \vec{\mathbf{F}}/l &= -1.30\widehat{\mathbf{k}}\,\text{N/m}.\end{aligned}$$

Significance
This large magnetic field creates a significant force on a small length of wire. As the angle of the magnetic field becomes more closely aligned to the current in the wire, there is less of a force on it, as seen from comparing parts a and b.

Note:
Exercise:

Problem:

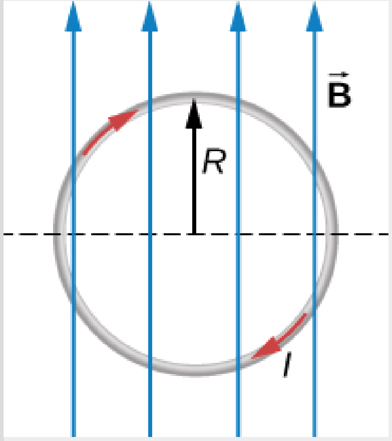
Check Your Understanding A straight, flexible length of copper wire is immersed in a magnetic field that is directed into the page. (a) If the wire's current runs in the +*x*-direction, which way will the wire bend? (b) Which way will the wire bend if the current runs in the –*x*-direction?

Solution:

a. bends upward; b. bends downward

Example:
Force on a Circular Wire
A circular current loop of radius *R* carrying a current *I* is placed in the *xy*-plane. A constant uniform magnetic field cuts through the loop parallel to the *y*-axis

([link]). Find the magnetic force on the upper half of the loop, the lower half of the loop, and the total force on the loop.



A loop of wire carrying a current in a magnetic field.

Strategy

The magnetic force on the upper loop should be written in terms of the differential force acting on each segment of the loop. If we integrate over each differential piece, we solve for the overall force on that section of the loop. The force on the lower loop is found in a similar manner, and the total force is the addition of these two forces.

Solution

A differential force on an arbitrary piece of wire located on the upper ring is:

Equation:

$$dF = IB\sin\theta dl.$$

where θ is the angle between the magnetic field direction ($+y$) and the segment of wire. A differential segment is located at the same radius, so using an arc-length formula, we have:

Equation:

$$\begin{array}{ccc} dl & = & Rd\theta \\ dF & = & IBR\sin\theta d\theta. \end{array}$$

In order to find the force on a segment, we integrate over the upper half of the circle, from 0 to π. This results in:

Equation:

$$F = IBR\int_{0}^{\pi} \sin\theta\, d\theta = IBR(-\cos\pi + \cos 0) = 2IBR.$$

The lower half of the loop is integrated from π to zero, giving us:

Equation:

$$F = IBR\int_{\pi}^{0} \sin\theta\, d\theta = IBR(-\cos 0 + \cos\pi) = -2IBR.$$

The net force is the sum of these forces, which is zero.

Significance

The total force on any closed loop in a uniform magnetic field is zero. Even though each piece of the loop has a force acting on it, the net force on the system is zero. (Note that there is a net torque on the loop, which we consider in the next section.)

Summary

- An electrical current produces a magnetic field around the wire.
- The directionality of the magnetic field produced is determined by the right hand rule-2, where your thumb points in the direction of the current and your fingers wrap around the wire in the direction of the magnetic field.
- The magnetic force on current-carrying conductors is given by $\overrightarrow{\mathbf{F}} = I\overrightarrow{\mathbf{l}} \times \overrightarrow{\mathbf{B}}$ where *I* is the current and *l* is the length of a wire in a uniform magnetic field *B*.

Conceptual Questions

Exercise:

Problem:

Describe the error that results from accidently using your left rather than your right hand when determining the direction of a magnetic force.

Exercise:

Problem:

Considering the magnetic force law, are the velocity and magnetic field always perpendicular? Are the force and velocity always perpendicular? What about the force and magnetic field?

Solution:

Velocity and magnetic field can be set together in any direction. If there is a force, the velocity is perpendicular to it. The magnetic field is also perpendicular to the force if it exists.

Exercise:

Problem:

Why can a nearby magnet distort a cathode ray tube television picture?

Exercise:

Problem:

A magnetic field exerts a force on the moving electrons in a current carrying wire. What exerts the force on a wire?

Solution:

A force on a wire is exerted by an external magnetic field created by a wire or another magnet.

Exercise:

Problem:

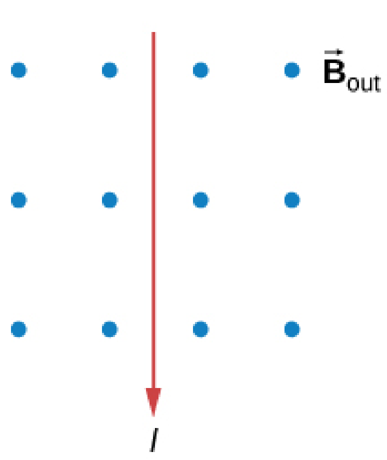
There are regions where the magnetic field of earth is almost perpendicular to the surface of Earth. What difficulty does this cause in the use of a compass?

Problems

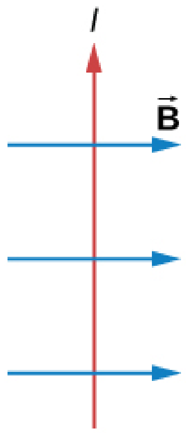
Exercise:

Problem:

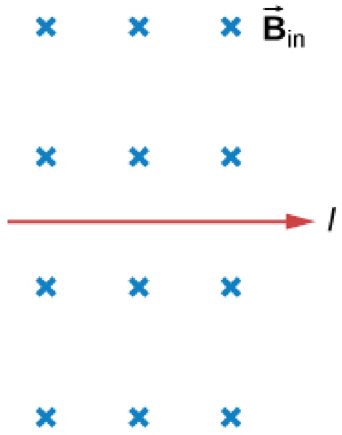
What is the direction of the magnetic force on the current in each of the six cases?


$\vec{\mathbf{B}}_{\text{out}}$
I

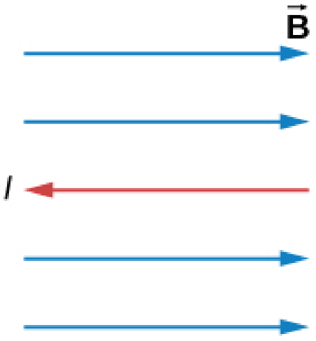
(a)


I
$\vec{\mathbf{B}}$

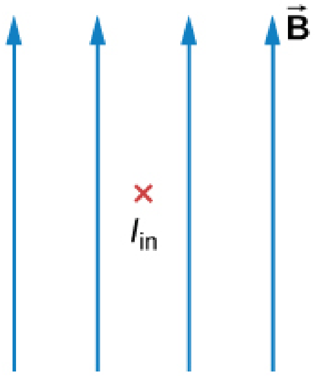
(b)


$\vec{\mathbf{B}}_{\text{in}}$
I

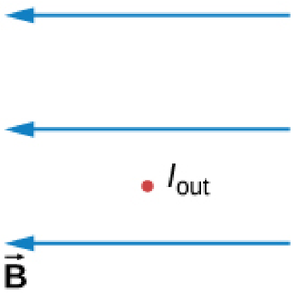
(c)


$\vec{\mathbf{B}}$
I

(d)


$\vec{\mathbf{B}}$
I_{in}

(e)


I_{out}
$\vec{\mathbf{B}}$

(f)

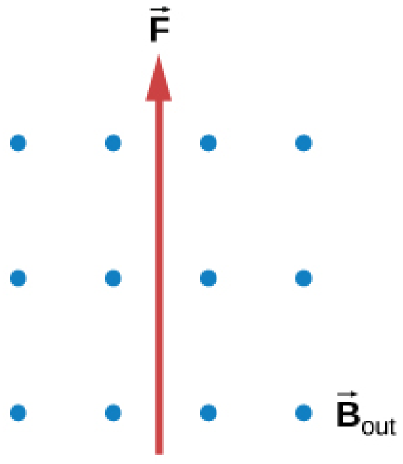
Solution:

a. left; b. into the page; c. up; d. no force; e. right; f. down

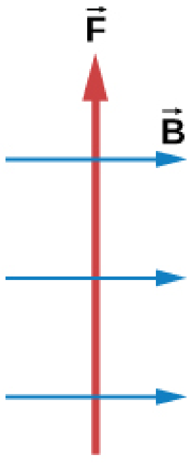
Exercise:

Problem:

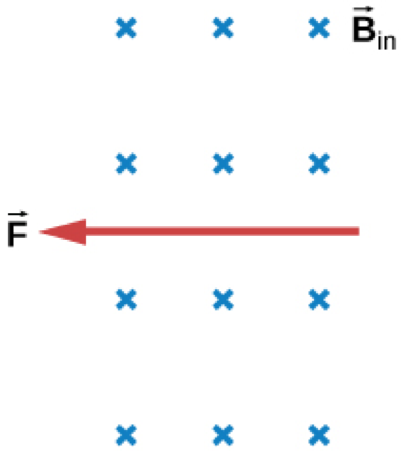
What is the direction of a current that experiences the magnetic force shown in each of the three cases, assuming the current runs perpendicular to $\vec{\mathbf{B}}$?


$\vec{F}$
$\vec{B}_{out}$

(a)



(b)

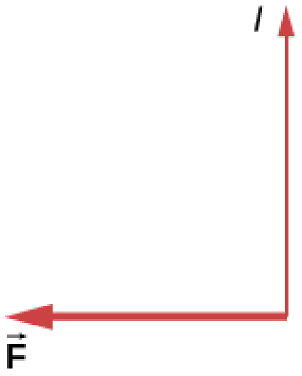

$\vec{B}_{in}$
$\vec{F}$

(c)

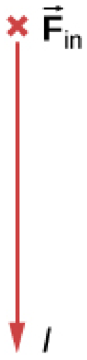
Exercise:

Problem:

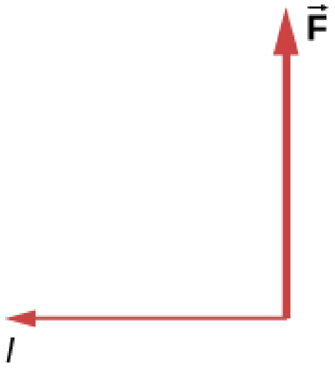
What is the direction of the magnetic field that produces the magnetic force shown on the currents in each of the three cases, assuming $\vec{\mathbf{B}}$ is perpendicular to I?


I
$\vec{\mathbf{F}}$

(a)


$\vec{\mathbf{F}}_{\text{in}}$
I

(b)


$\vec{\mathbf{F}}$
I

(c)

Solution:

a. into the page; b. left; c. out of the page

Exercise:

Problem:

(a) What is the force per meter on a lightning bolt at the equator that carries 20,000 A perpendicular to Earth’s $3.0 \times 10^{-5}\text{T}$ field? (b) What is the direction of the force if the current is straight up and Earth’s field direction is due north, parallel to the ground?

Exercise:

Problem:

(a) A dc power line for a light-rail system carries 1000 A at an angle of 30.0º to Earth’s $5.0 \times 10^{-5}\text{T}$ field. What is the force on a 100-m section of this line? (b) Discuss practical concerns this presents, if any.

Solution:

a. 2.50 N; b. This means that the light-rail power lines must be attached in order not to be moved by the force caused by Earth’s magnetic field.

Exercise:

Problem:

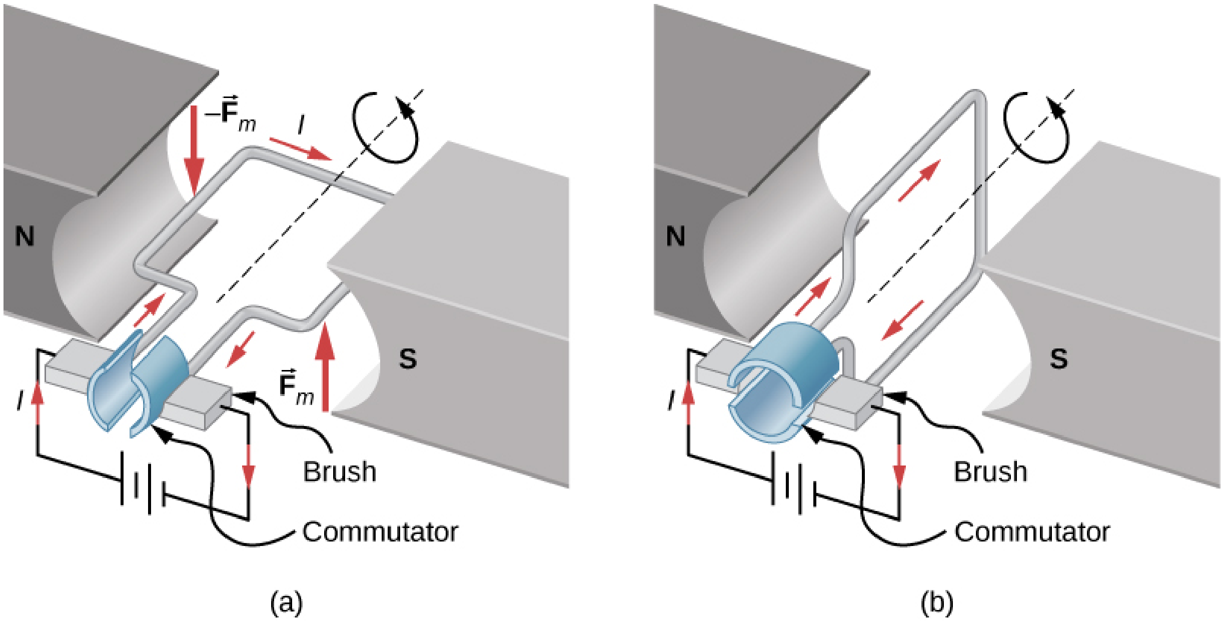
A wire carrying a 30.0-A current passes between the poles of a strong magnet that is perpendicular to its field and experiences a 2.16-N force on the 4.00 cm of wire in the field. What is the average field strength?

Force and Torque on a Current Loop

By the end of this section, you will be able to:

- Evaluate the net force on a current loop in an external magnetic field
- Evaluate the net torque on a current loop in an external magnetic field
- Define the magnetic dipole moment of a current loop

Motors are the most common application of magnetic force on current-carrying wires. Motors contain loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft. Electrical energy is converted into mechanical work in the process. Once the loop's surface area is aligned with the magnetic field, the direction of current is reversed, so there is a continual torque on the loop ([link]). This reversal of the current is done with commutators and brushes. The commutator is set to reverse the current flow at set points to keep continual motion in the motor. A basic commutator has three contact areas to avoid and dead spots where the loop would have zero instantaneous torque at that point. The brushes press against the commutator, creating electrical contact between parts of the commutator during the spinning motion.



A simplified version of a dc electric motor. (a) The rectangular wire loop is placed in a magnetic field. The forces on the wires closest to the magnetic poles (N and S) are opposite in direction as determined by the right-hand rule-1. Therefore, the loop has a net torque and rotates to the position shown in (b). (b) The brushes now touch the commutator segments so that no current flows through the loop. No torque acts on the loop, but the loop continues to spin from the initial velocity given to it in part (a). By the time the loop flips over, current flows through the wires again but now in the opposite direction, and the process repeats as in part (a). This causes continual rotation of the loop.

In a uniform magnetic field, a current-carrying loop of wire, such as a loop in a motor, experiences both forces and torques on the loop. [link] shows a rectangular loop of wire that carries a current *I* and has sides of lengths *a* and *b*. The loop is in a uniform magnetic field: $\vec{\mathbf{B}} = B\hat{\mathbf{j}}$. The magnetic force on a straight current-carrying wire of length *l* is given by $I\vec{\mathbf{l}} \times \vec{\mathbf{B}}$. To find the net force on the loop, we have to apply this equation to each of the four sides. The force on side 1 is

Equation:

$$\vec{\mathbf{F}}_1 = IaB\sin(90^\circ - \theta)\hat{\mathbf{i}} = IaB\cos\theta\hat{\mathbf{i}}$$

where the direction has been determined with the RHR-1. The current in side 3 flows in the opposite direction to that of side 1, so

Equation:

$$\vec{\mathbf{F}}_3 = -IaB\sin(90^\circ + \theta)\hat{\mathbf{i}} = -IaB\cos\theta\hat{\mathbf{i}}.$$

The currents in sides 2 and 4 are perpendicular to $\vec{\mathbf{B}}$ and the forces on these sides are

Equation:

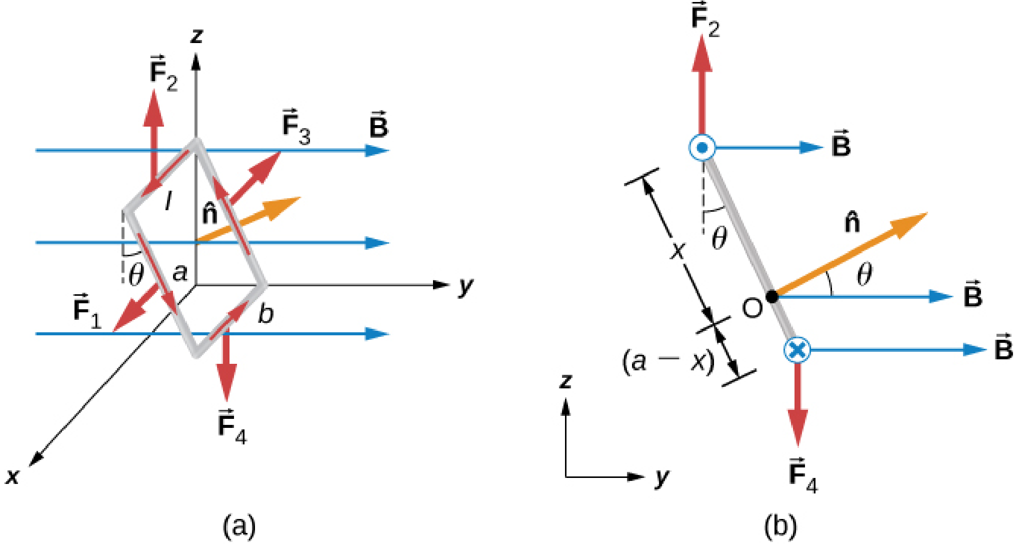
$$\vec{\mathbf{F}}_2 = IbB\widehat{\mathbf{k}}, \;\; \vec{\mathbf{F}}_4 = -IbB\widehat{\mathbf{k}}.$$

We can now find the net force on the loop:

Equation:

$$\sum \vec{\mathbf{F}}_{\text{net}} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 + \vec{\mathbf{F}}_3 + \vec{\mathbf{F}}_4 = 0.$$

Although this result $(\Sigma F = 0)$ has been obtained for a rectangular loop, it is far more general and holds for current-carrying loops of arbitrary shapes; that is, there is no net force on a current loop in a uniform magnetic field.



(a) A rectangular current loop in a uniform magnetic field is subjected to a net torque but not a net force. (b) A side view of the coil.

To find the net torque on the current loop shown in [link], we first consider F_1 and F_3. Since they have the same line of action and are equal and opposite, the sum of their torques about any axis is zero (see Fixed-Axis Rotation). Thus, if there is any torque on the loop, it must be furnished by F_2 and F_4. Let's calculate the torques around the axis that passes through point *O* of [link] (a side view of the coil) and is perpendicular to the plane of the page. The point *O* is a distance *x* from side 2 and a distance $(a-x)$ from side 4 of the loop. The moment arms of F_2 and F_4 are $x\sin\theta$ and $(a-x)\sin\theta$, respectively, so the net torque on the loop is

Equation:

$$\begin{aligned}\sum\vec{\boldsymbol{\tau}}=\vec{\boldsymbol{\tau}}_1+\vec{\boldsymbol{\tau}}_2+\vec{\boldsymbol{\tau}}_3+\vec{\boldsymbol{\tau}}_4 &= F_2x\sin\theta\hat{\mathbf{i}}-F_4(a-x)\sin(\theta)\hat{\mathbf{i}}\\ &= -IbBx\sin\theta\hat{\mathbf{i}}-IbB(a-x)\sin\theta\hat{\mathbf{i}}.\end{aligned}$$

This simplifies to

Equation:

$$\vec{\boldsymbol{\tau}}=-IAB\sin\theta\hat{\mathbf{i}}$$

where $A=ab$ is the area of the loop.

Notice that this torque is independent of *x*; it is therefore independent of where point *O* is located in the plane of the current loop. Consequently, the loop experiences the same torque from the magnetic field about any axis in the plane of the loop and parallel to the *x*-axis.

A closed-current loop is commonly referred to as a **magnetic dipole** and the term *IA* is known as its **magnetic dipole moment** μ. Actually, the magnetic dipole moment is a vector that is defined as

Equation:

$$\vec{\boldsymbol{\mu}}=IA\hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is a unit vector directed perpendicular to the plane of the loop (see [link]). The direction of $\hat{\mathbf{n}}$ is obtained with the RHR-2—if you curl the fingers of your right hand in the direction of current flow in the loop, then your thumb points along $\hat{\mathbf{n}}$. If the loop contains *N* turns of wire, then its magnetic dipole moment is given by

Note:

Equation:

$$\vec{\boldsymbol{\mu}}=NIA\hat{\mathbf{n}}.$$

In terms of the magnetic dipole moment, the torque on a current loop due to a uniform magnetic field can be written simply as

Note:
Equation:

$$\vec{\tau} = \vec{\mu} \times \vec{\mathbf{B}}.$$

This equation holds for a current loop in a two-dimensional plane of arbitrary shape.

Using a calculation analogous to that found in Capacitance for an electric dipole, the potential energy of a magnetic dipole is

Note:
Equation:

$$U = -\vec{\mu} \cdot \vec{\mathbf{B}}.$$

Example:
Forces and Torques on Current-Carrying Loops
A circular current loop of radius 2.0 cm carries a current of 2.0 mA. (a) What is the magnitude of its magnetic dipole moment? (b) If the dipole is oriented at 30 degrees to a uniform magnetic field of magnitude 0.50 T, what is the magnitude of the torque it experiences and what is its potential energy?
Strategy
The dipole moment is defined by the current times the area of the loop. The area of the loop can be calculated from the area of the circle. The torque on the loop and potential energy are calculated from identifying the magnetic moment, magnetic field, and angle oriented in the field.
Solution

a. The magnetic moment μ is calculated by the current times the area of the loop or πr^2.
   Equation:

$$\mu = IA = (2.0 \times 10^{-3}\text{ A})(\pi(0.02\text{ m})^2) = 2.5 \times 10^{-6}\text{ A} \cdot \text{m}^2$$

b. The torque and potential energy are calculated by identifying the magnetic moment, magnetic field, and the angle between these two vectors. The calculations of these quantities are:
   Equation:

$$\begin{aligned} \tau &= \vec{\mu} \times \vec{\mathbf{B}} = \mu B \sin\theta = \left(2.5 \times 10^{-6}\text{A}\cdot\text{m}^2\right)(0.50\text{T})\sin(30^\circ) = 6.3 \times 10^{-7}\text{N}\cdot\text{m} \\ U &= -\vec{\mu}\cdot\vec{\mathbf{B}} = -\mu B\cos\theta = -\left(2.5 \times 10^{-6}\text{A}\cdot\text{m}^2\right)(0.50\text{T})\cos(30^\circ) = -1.1 \times 10^{-6}\text{J}. \end{aligned}$$

Significance
The concept of magnetic moment at the atomic level is discussed in the next chapter. The concept of aligning the magnetic moment with the magnetic field is the functionality of devices like magnetic motors, whereby switching the external magnetic field results in a constant spinning of the loop as it tries to align with the field to minimize its potential energy.

Note:
Exercise:

Problem: Check Your Understanding

In what orientation would a magnetic dipole have to be to produce (a) a maximum torque in a magnetic field? (b) A maximum energy of the dipole?

Solution:

a. aligned or anti-aligned; b. perpendicular

Summary

- The net force on a current-carrying loop of any plane shape in a uniform magnetic field is zero.
- The net torque τ on a current-carrying loop of any shape in a uniform magnetic field is calculated using $\tau = \vec{\mu} \times \vec{\mathbf{B}}$ where $\vec{\mu}$ is the magnetic dipole moment and $\vec{\mathbf{B}}$ is the magnetic field strength.
- The magnetic dipole moment μ is the product of the number of turns of wire N, the current in the loop I, and the area of the loop A or $\vec{\mu} = NIA\hat{\mathbf{n}}$.

Problems

Exercise:

Problem:

(a) By how many percent is the torque of a motor decreased if its permanent magnets lose 5.0% of their strength? (b) How many percent would the current need to be increased to return the torque to original values?

Solution:

a. $\tau = NIAB$, so τ decreases by 5.00% if B decreases by 5.00%; b. 5.26% increase

Exercise:

Problem:

(a) What is the maximum torque on a 150-turn square loop of wire 18.0 cm on a side that carries a 50.0-A current in a 1.60-T field? (b) What is the torque when θ is 10.9º?

Exercise:

Problem:

Find the current through a loop needed to create a maximum torque of $9.0 \text{ N} \cdot \text{m}$. The loop has 50 square turns that are 15.0 cm on a side and is in a uniform 0.800-T magnetic field.

Solution:

10.0 A

Exercise:

Problem:

Calculate the magnetic field strength needed on a 200-turn square loop 20.0 cm on a side to create a maximum torque of 300 N · m if the loop is carrying 25.0 A.

Exercise:

Problem:

Since the equation for torque on a current-carrying loop is $\tau = NIAB \sin \theta$, the units of N · m must equal units of A · m^2 T. Verify this.

Solution:

$$\text{A} \cdot \text{m}^2 \cdot \text{T} = \text{A} \cdot \text{m}^2 . \frac{\text{N}}{\text{A} \cdot \text{m}} = \text{N} \cdot \text{m}$$

Exercise:

Problem:

(a) At what angle θ is the torque on a current loop 90.0% of maximum? (b) 50.0% of maximum? (c) 10.0% of maximum?

Exercise:

Problem:

A proton has a magnetic field due to its spin. The field is similar to that created by a circular current loop $0.65 \times 10^{-15} \text{m}$ in radius with a current of $1.05 \times 10^{4} \text{A}$. Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

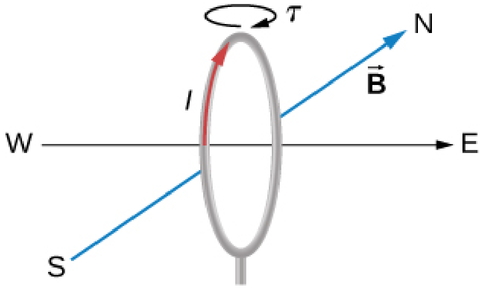
Solution:

$3.48 \times 10^{-26} \text{ N} \cdot \text{m}$

Exercise:

Problem:

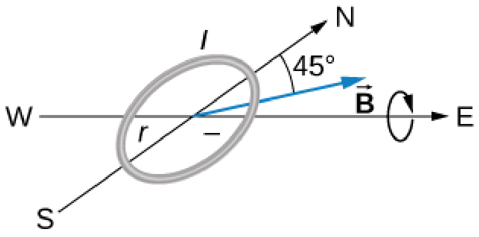
(a) A 200-turn circular loop of radius 50.0 cm is vertical, with its axis on an east-west line. A current of 100 A circulates clockwise in the loop when viewed from the east. Earth’s field here is due north, parallel to the ground, with a strength of $3.0 \times 10^{-5}\mathrm{T}$. What are the direction and magnitude of the torque on the loop? (b) Does this device have any practical applications as a motor?



Exercise:

Problem:

Repeat the previous problem, but with the loop lying flat on the ground with its current circulating counterclockwise (when viewed from above) in a location where Earth’s field is north, but at an angle 45.0° below the horizontal and with a strength of $6.0 \times 10^{-5}\mathrm{T}$.



Solution:

$0.666\ \mathrm{N} \cdot \mathrm{m}$

Glossary

magnetic dipole

closed-current loop

magnetic dipole moment

term *IA* of the magnetic dipole, also called μ

motor (dc)

loop of wire in a magnetic field; when current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft; electrical energy is converted into mechanical work in the process

The Hall Effect

By the end of this section, you will be able to:

- Explain a scenario where the magnetic and electric fields are crossed and their forces balance each other as a charged particle moves through a velocity selector
- Compare how charge carriers move in a conductive material and explain how this relates to the Hall effect

In 1879, E.H. Hall devised an experiment that can be used to identify the sign of the predominant charge carriers in a conducting material. From a historical perspective, this experiment was the first to demonstrate that the charge carriers in most metals are negative.

Note:
Visit this website to find more information about the Hall effect.

We investigate the **Hall effect** by studying the motion of the free electrons along a metallic strip of width *l* in a constant magnetic field ([link]). The electrons are moving from left to right, so the magnetic force they experience pushes them to the bottom edge of the strip. This leaves an excess of positive charge at the top edge of the strip, resulting in an electric field *E* directed from top to bottom. The charge concentration at both edges builds up until the electric force on the electrons in one direction is balanced by the magnetic force on them in the opposite direction. Equilibrium is reached when:

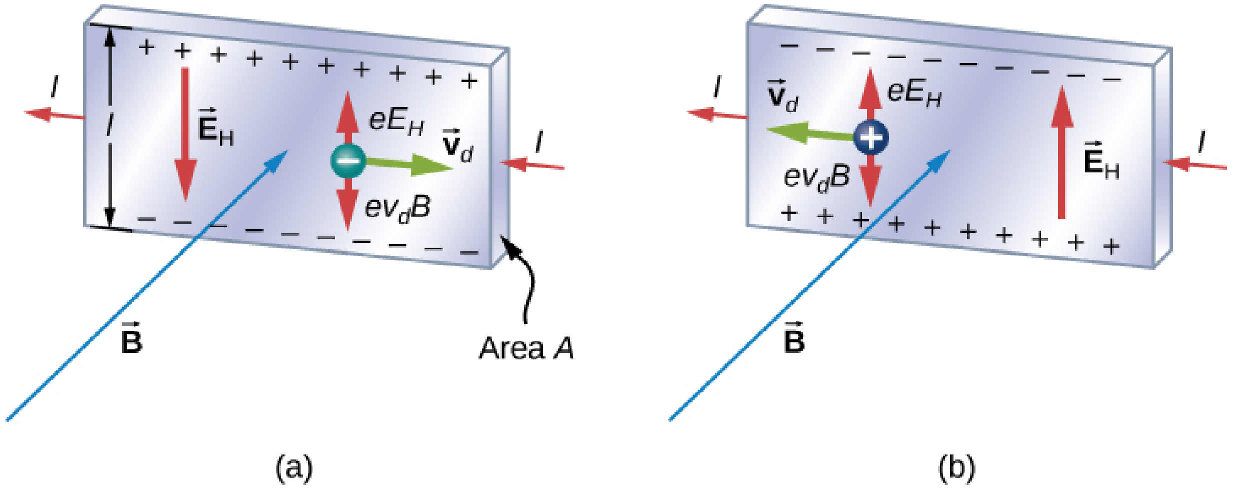
Equation:

$$eE = ev_d B$$

where e is the magnitude of the electron charge, v_d is the drift speed of the electrons, and E is the magnitude of the electric field created by the separated charge. Solving this for the drift speed results in

Note:
Equation:

$$v_d = \frac{E}{B}.$$



In the Hall effect, a potential difference between the top and bottom edges of the metal strip is produced when moving charge carriers are deflected by the magnetic field. (a) Hall effect for negative charge carriers; (b) Hall effect for positive charge carriers.

A scenario where the electric and magnetic fields are perpendicular to one another is called a crossed-field situation. If these fields produce equal and opposite forces on a charged particle with the velocity that equates the forces, these particles are able to pass through an apparatus, called a **velocity selector**, undeflected. This velocity is represented in [link]. Any other velocity of a charged particle sent into the same fields would be deflected by the magnetic force or electric force.

Going back to the Hall effect, if the current in the strip is *I*, then from Current and Resistance, we know that
Equation:

$$I = nev_d A$$

where *n* is the number of charge carriers per volume and *A* is the cross-sectional area of the strip. Combining the equations for v_d and *I* results in
Equation:

$$I = ne\left(\frac{E}{B}\right)A.$$

The field *E* is related to the potential difference *V* between the edges of the strip by
Equation:

$$E = \frac{V}{l}.$$

The quantity V is called the Hall potential and can be measured with a voltmeter. Finally, combining the equations for I and E gives us

Note:
Equation:

$$V = \frac{IBl}{neA}$$

where the upper edge of the strip in [link] is positive with respect to the lower edge.

We can also combine [link] and [link] to get an expression for the Hall voltage in terms of the magnetic field:

Note:
Equation:

$$V = Blv_d.$$

What if the charge carriers are positive, as in [link]? For the same current I, the magnitude of V is still given by [link]. However, the upper edge is now negative with respect to the lower edge. Therefore, by simply measuring the sign of V, we can determine the sign of the majority charge carriers in a metal.

Hall potential measurements show that electrons are the dominant charge carriers in most metals. However, Hall potentials indicate that for a few metals, such as tungsten, beryllium, and many semiconductors, the majority of charge carriers are positive. It turns out that conduction by positive charge is caused by the migration of missing electron sites (called holes) on ions. Conduction by holes is studied later in Condensed Matter Physics.

The Hall effect can be used to measure magnetic fields. If a material with a known density of charge carriers n is placed in a magnetic field and V is measured, then the field can be determined from [link]. In research laboratories where the fields of electromagnets used for precise measurements have to be extremely steady, a “Hall probe” is commonly used as part of an electronic circuit that regulates the field.

Example:

Velocity Selector

An electron beam enters a crossed-field velocity selector with magnetic and electric fields of 2.0 mT and $6.0 \times 10^3 \text{ N/C}$, respectively. (a) What must the velocity of the electron beam be to traverse the crossed fields undeflected? If the electric field is turned off, (b) what is the acceleration of the electron beam and (c) what is the radius of the circular motion that results?

Strategy

The electron beam is not deflected by either of the magnetic or electric fields if these forces are balanced. Based on these balanced forces, we calculate the velocity of the beam. Without the electric field, only the magnetic force is used in Newton’s second law to find the acceleration. Lastly, the radius of the path is based on the resulting circular motion from the magnetic force.

Solution

a. The velocity of the unperturbed beam of electrons with crossed fields is calculated by [link]:
   Equation:

$$v_d = \frac{E}{B} = \frac{6 \times 10^3 \text{ N/C}}{2 \times 10^{-3} \text{ T}} = 3 \times 10^6 \text{m/s}.$$

b. The acceleration is calculated from the net force from the magnetic field, equal to mass times acceleration. The magnitude of the acceleration is:
   Equation:

$$\begin{aligned} ma &= qvB \\ a &= \frac{qvB}{m} = \frac{(1.6 \times 10^{-19}\text{C})(3 \times 10^6 \text{ m/s})(2 \times 10^{-3}\text{T})}{9.1 \times 10^{-31}\text{kg}} = 1.1 \times 10^{15} \text{ m/s}^2. \end{aligned}$$

c. The radius of the path comes from a balance of the circular and magnetic forces, or [link]:
   Equation:

$$r = \frac{mv}{qB} = \frac{(9.1 \times 10^{-31}\text{kg})(3 \times 10^6 \text{ m/s})}{(1.6 \times 10^{-19}\text{C})(2 \times 10^{-3}\text{T})} = 8.5 \times 10^{-3}\text{m}.$$

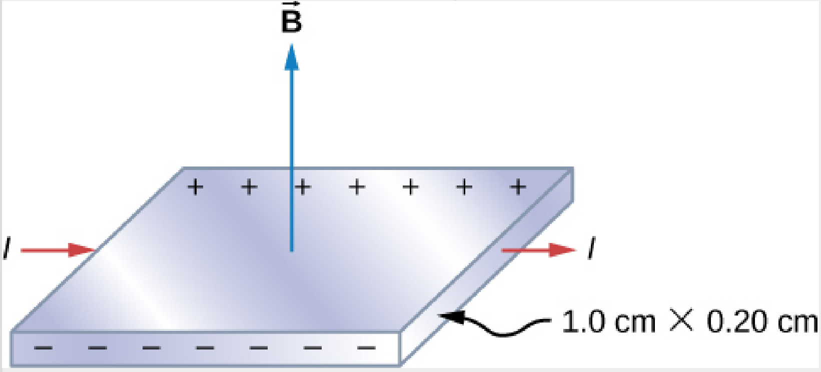
Significance

If electrons in the beam had velocities above or below the answer in part (a), those electrons would have a stronger net force exerted by either the magnetic or electric field. Therefore, only those electrons at this specific velocity would make it through.

Example:

The Hall Potential in a Silver Ribbon

[link] shows a silver ribbon whose cross section is 1.0 cm by 0.20 cm. The ribbon carries a current of 100 A from left to right, and it lies in a uniform magnetic field of magnitude 1.5 T. Using a density value of $n = 5.9 \times 10^{28}$ electrons per cubic meter for silver, find the Hall potential between the edges of the ribbon.



Finding the Hall potential in a silver ribbon in a magnetic field is shown.

Strategy

Since the majority of charge carriers are electrons, the polarity of the Hall voltage is that indicated in the figure. The value of the Hall voltage is calculated using [link]:

Equation:

$$V = \frac{IBl}{neA}.$$

Solution

When calculating the Hall voltage, we need to know the current through the material, the magnetic field, the length, the number of charge carriers, and the area. Since all of these are given, the Hall voltage is calculated as:

Equation:

$$V = \frac{IBl}{neA} = \frac{(100\text{ A})(1.5\text{ T})\left(1.0 \times 10^{-2}\text{m}\right)}{\left(5.9 \times 10^{28}/\text{m}^3\right)\left(1.6 \times 10^{-19}\text{C}\right)\left(2.0 \times 10^{-5}\text{m}^2\right)} = 7.9 \times 10^{-6}\text{V}.$$

Significance
As in this example, the Hall potential is generally very small, and careful experimentation with sensitive equipment is required for its measurement.

Note:
Exercise:

Problem:

Check Your Understanding A Hall probe consists of a copper strip, $n = 8.5 \times 10^{28}$ electrons per cubic meter, which is 2.0 cm wide and 0.10 cm thick. What is the magnetic field when *I* = 50 A and the Hall potential is (a) $4.0 \mu \text{V}$ and (b) $6.0 \mu \text{V}$?

Solution:

a. 1.1 T; b. 1.6 T

Summary

- Perpendicular electric and magnetic fields exert equal and opposite forces for a specific velocity of entering particles, thereby acting as a velocity selector. The velocity that passes through undeflected is calculated by $v = \frac{E}{B}$.
- The Hall effect can be used to measure the sign of the majority of charge carriers for metals. It can also be used to measure a magnetic field.

Conceptual Questions

Exercise:

Problem:

Hall potentials are much larger for poor conductors than for good conductors. Why?

Solution:

Poor conductors have a lower charge carrier density, *n*, which, based on the Hall effect formula, relates to a higher Hall potential. Good conductors have a higher charge carrier density, thereby a lower Hall potential.

Problems

Exercise:

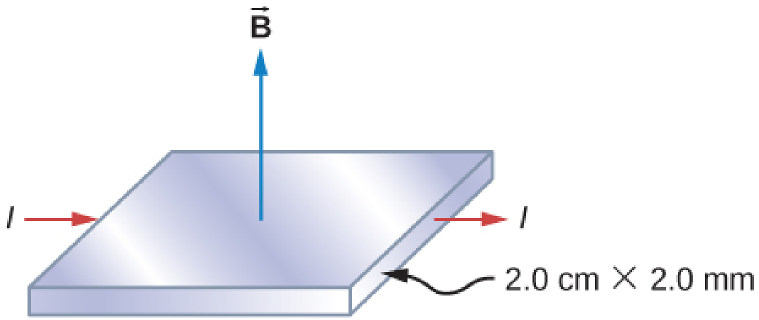
Problem:

A strip of copper is placed in a uniform magnetic field of magnitude 2.5 T. The Hall electric field is measured to be $1.5 \times 10^{-3} \mathrm{V/m}$. (a) What is the drift speed of the conduction electrons? (b) Assuming that n = 8.0×10^{28} electrons per cubic meter and that the cross-sectional area of the strip is $5.0 \times 10^{-6} \mathrm{m}^2$, calculate the current in the strip. (c) What is the Hall coefficient 1/nq?

Exercise:

Problem:

The cross-sectional dimensions of the copper strip shown are 2.0 cm by 2.0 mm. The strip carries a current of 100 A, and it is placed in a magnetic field of magnitude B = 1.5 T. What are the value and polarity of the Hall potential in the copper strip?



Solution:

$5.8 \times 10^{-6} \mathrm{V}$

Exercise:

Problem:

The magnitudes of the electric and magnetic fields in a velocity selector are $1.8 \times 10^{5} \mathrm{V/m}$ and 0.080 T, respectively. (a) What speed must a proton have to pass through the selector? (b) Also calculate the speeds required for an alpha-particle and a singly ionized $_{s}O^{16}$ atom to pass through the selector.

Exercise:

Problem:

A charged particle moves through a velocity selector at constant velocity. In the selector, $E = 1.0 \times 10^4 \mathrm{N/C}$ and $B = 0.250$ T. When the electric field is turned off, the charged particle travels in a circular path of radius 3.33 mm. Determine the charge-to-mass ratio of the particle.

Solution:

$4.8 \times 10^7 \mathrm{C/kg}$

Exercise:

Problem:

A Hall probe gives a reading of $1.5 \mu\mathrm{V}$ for a current of 2 A when it is placed in a magnetic field of 1 T. What is the magnetic field in a region where the reading is $2 \mu\mathrm{V}$ for 1.7 A of current?

Glossary

Hall effect	creation of voltage across a current-carrying conductor by a magnetic field
velocity selector	apparatus where the crossed electric and magnetic fields produce equal and opposite forces on a charged particle moving with a specific velocity; this particle moves through the velocity selector not affected by either field while particles moving with different velocities are deflected by the apparatus

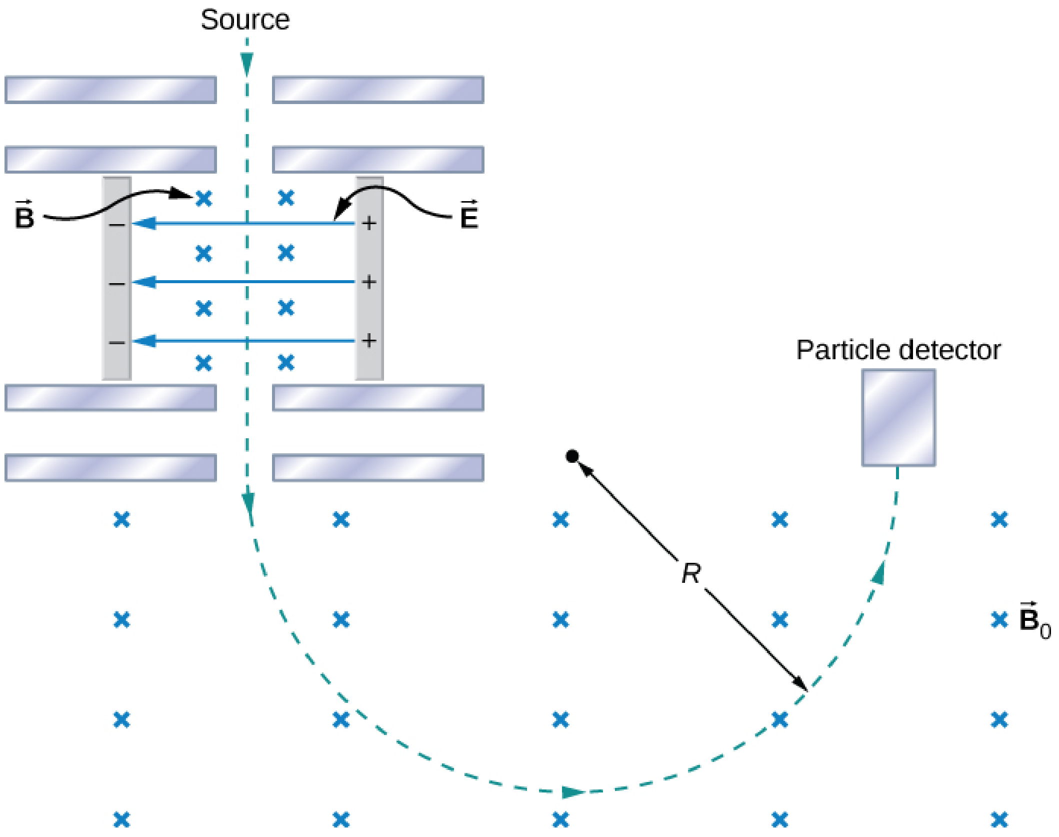
Applications of Magnetic Forces and Fields
By the end of this section, you will be able to:

- Explain how a mass spectrometer works to separate charges
- Explain how a cyclotron works

Being able to manipulate and sort charged particles allows deeper experimentation to understand what matter is made of. We first look at a mass spectrometer to see how we can separate ions by their charge-to-mass ratio. Then we discuss cyclotrons as a method to accelerate charges to very high energies.

Mass Spectrometer

The **mass spectrometer** is a device that separates ions according to their charge-to-mass ratios. One particular version, the Bainbridge mass spectrometer, is illustrated in [link]. Ions produced at a source are first sent through a velocity selector, where the magnetic force is equally balanced with the electric force. These ions all emerge with the same speed $v=E/B$ since any ion with a different velocity is deflected preferentially by either the electric or magnetic force, and ultimately blocked from the next stage. They then enter a uniform magnetic field B_0 where they travel in a circular path whose radius R is given by [link]. The radius is measured by a particle detector located as shown in the figure.



A schematic of the Bainbridge mass spectrometer, showing charged particles leaving a source, followed by a velocity selector where the electric and magnetic forces are balanced, followed by a region of uniform magnetic field where the particle is ultimately detected.

The relationship between the charge-to-mass ratio q/m and the radius R is determined by combining [link] and [link]:

Note:
Equation:

$$\frac{q}{m} = \frac{E}{BB_0R}.$$

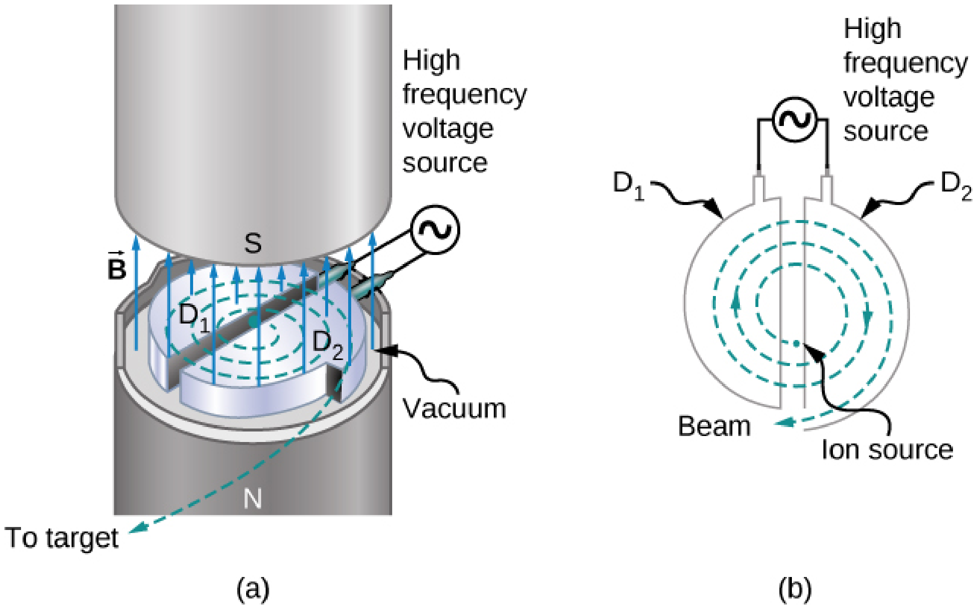
Since most ions are singly charged $\left(q = 1.6 \times 10^{-19}\ \text{C}\right)$, measured values of R can be used with this equation to determine the mass of ions. With modern instruments, masses can be determined to one part in 10^8.

An interesting use of a spectrometer is as part of a system for detecting very small leaks in a research apparatus. In low-temperature physics laboratories, a device known as a dilution refrigerator uses a mixture of He-3, He-4, and other cryogens to reach temperatures well below 1 K. The performance of the refrigerator is severely hampered if even a minute leak between its various components occurs. Consequently, before it is cooled down to the desired temperature, the refrigerator is subjected to a leak test. A small quantity of gaseous helium is injected into one of its compartments, while an adjacent, but supposedly isolated, compartment is connected to a high-vacuum pump to which a mass spectrometer is attached. A heated filament ionizes any helium atoms evacuated by the pump. The detection of these ions by the spectrometer then indicates a leak between the two compartments of the dilution refrigerator.

In conjunction with gas chromatography, mass spectrometers are used widely to identify unknown substances. While the gas chromatography portion breaks down the substance, the mass spectrometer separates the resulting ionized molecules. This technique is used with fire debris to ascertain the cause, in law enforcement to identify illegal drugs, in security to identify explosives, and in many medicinal applications.

Cyclotron

The **cyclotron** was developed by E.O. Lawrence to accelerate charged particles (usually protons, deuterons, or alpha-particles) to large kinetic energies. These particles are then used for nuclear-collision experiments to produce radioactive isotopes. A cyclotron is illustrated in [link]. The particles move between two flat, semi-cylindrical metallic containers D1 and D2, called **dees**. The dees are enclosed in a larger metal container, and the apparatus is placed between the poles of an electromagnet that provides a uniform magnetic field. Air is removed from the large container so that the particles neither lose energy nor are deflected because of collisions with air molecules. The dees are connected to a high-frequency voltage source that provides an alternating electric field in the small region between them. Because the dees are made of metal, their interiors are shielded from the electric field.


High frequency voltage source
S
$\vec{B}$
D_1
D_2
Vacuum
N
To target
(a)
High frequency voltage source
D_1
D_2
Beam
Ion source
(b)

The inside of a cyclotron. A uniform magnetic field is applied as circulating protons travel through the dees, gaining energy as they traverse through the gap between the dees.

Suppose a positively charged particle is injected into the gap between the dees when D2 is at a positive potential relative to D1. The particle is then accelerated across the gap and enters D1 after gaining kinetic energy qV, where V is the average potential difference the particle experiences between the dees. When the particle is inside D1, only the uniform magnetic field $\vec{\mathbf{B}}$ of the electromagnet acts on it, so the particle moves in a circle of radius

Equation:

$$r = \frac{mv}{qB}$$

with a period of

Equation:

$$T = \frac{2\pi m}{qB}.$$

The period of the alternating voltage course is set at T, so while the particle is inside D1, moving along its semicircular orbit in a time $T/2$, the polarity of the dees is reversed. When the particle reenters the gap, D1 is positive with respect to D2, and the particle is again accelerated across the gap, thereby gaining a kinetic energy qV. The particle then enters D2, circulates in a slightly larger circle, and emerges from D2 after spending a time $T/2$ in this dee. This process repeats until the orbit of the particle reaches the boundary of the dees. At that point, the particle (actually, a beam of particles) is extracted from the cyclotron and used for some experimental purpose.

The operation of the cyclotron depends on the fact that, in a uniform magnetic field, a particle's orbital period is independent of its radius and its kinetic energy. Consequently, the period of the alternating voltage source need only be set at the one value given by [link]. With that setting, the electric field accelerates particles every time they are between the dees.

If the maximum orbital radius in the cyclotron is R, then from [link], the maximum speed of a circulating particle of mass m and charge q is

Note:
Equation:

$$v_{\max} = \text{mfrac}.$$

Thus, its kinetic energy when ejected from the cyclotron is
Equation:

$$\frac{1}{2}m{v_{\max}}^2 = \frac{q^2B^2R^2}{2m}.$$

The maximum kinetic energy attainable with this type of cyclotron is approximately 30 MeV. Above this energy, relativistic effects become important, which causes the orbital period to increase with the radius. Up to energies of several hundred MeV, the relativistic effects can be compensated for by making the magnetic field gradually increase with the radius of the orbit. However, for higher energies, much more elaborate methods must be used to accelerate particles.

Particles are accelerated to very high energies with either linear accelerators or synchrotrons. The linear accelerator accelerates particles continuously with the electric field of an electromagnetic wave that travels down a long evacuated tube. The Stanford Linear Accelerator (SLAC) is about 3.3 km long and accelerates electrons and positrons (positively charged electrons) to energies of 50 GeV. The synchrotron is constructed so that its bending magnetic field increases with particle speed in such a way that the particles stay in an orbit of fixed radius. The world's highest-energy synchrotron is located at CERN, which is on the Swiss-French border near Geneva. CERN has been of recent interest with the verified discovery of the Higgs Boson (see Particle Physics and Cosmology). This synchrotron can accelerate beams of approximately 10^{13} protons to energies of about 10^3 GeV.

Example:

Accelerating Alpha-Particles in a Cyclotron

A cyclotron used to accelerate alpha-particles ($m = 6.64 \times 10^{-27}\text{kg}, q = 3.2 \times 10^{-19}\text{C}$) has a radius of 0.50 m and a magnetic field of 1.8 T. (a) What is the period of revolution of the alpha-particles? (b) What is their maximum kinetic energy?

Strategy

a. The period of revolution is approximately the distance traveled in a circle divided by the speed. Identifying that the magnetic force applied is the centripetal force, we can derive the period formula.
b. The kinetic energy can be found from the maximum speed of the beam, corresponding to the maximum radius within the cyclotron.

Solution

a. By identifying the mass, charge, and magnetic field in the problem, we can calculate the period:
   Equation:

$$T = \frac{2\pi m}{qB} = \frac{2\pi\left(6.64 \times 10^{-27}\text{kg}\right)}{\left(3.2 \times 10^{-19}\text{C}\right)(1.8\text{T})} = 7.3 \times 10^{-8}\text{s}.$$

b. By identifying the charge, magnetic field, radius of path, and the mass, we can calculate the maximum kinetic energy:
   Equation:

$$\frac{1}{2}m{v_{\text{max}}}^2 = \frac{q^2B^2R^2}{2m} = \frac{\left(3.2 \times 10^{-19}\text{C}\right)^2(1.8\text{T})^2(0.50\text{m})^2}{2(6.65 \times 10^{-27}\text{kg})} = 6.2 \times 10^{-12}\text{J} = 39\text{MeV}.$$

Note:

Exercise:

Problem:

Check Your Understanding A cyclotron is to be designed to accelerate protons to kinetic energies of 20 MeV using a magnetic field of 2.0 T. What is the required radius of the cyclotron?

Solution:

0.32 m

Summary

- A mass spectrometer is a device that separates ions according to their charge-to-mass ratios by first sending them through a velocity selector, then a uniform magnetic field.
- Cyclotrons are used to accelerate charged particles to large kinetic energies through applied electric and magnetic fields.

Key Equations

Force on a charge in a magnetic field	$\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$
Magnitude of magnetic force	$F = qvB\sin\theta$
Radius of a particle's path in a magnetic field	$r = \frac{mv}{qB}$
Period of a particle's motion in a magnetic field	$T = \frac{2\pi m}{qB}$
Force on a current-carrying wire in a uniform magnetic field	$\overrightarrow{\mathbf{F}} = I\overrightarrow{\mathbf{l}} \times \overrightarrow{\mathbf{B}}$
Magnetic dipole moment	$\vec{\mu} = NIA\hat{\mathbf{n}}$
Torque on a current loop	$\vec{\tau} = \vec{\mu} \times \vec{\mathbf{B}}$
Energy of a magnetic dipole	$U = -\vec{\mu} \cdot \vec{\mathbf{B}}$
Drift velocity in crossed electric and magnetic fields	$v_d = \frac{E}{B}$
Hall potential	$V = \frac{IBl}{neA}$
Hall potential in terms of drift velocity	$V = Blv_d$
Charge-to-mass ratio in a mass spectrometer	$\frac{q}{m} = \frac{E}{BB_0R}$
Maximum speed of a particle in a cyclotron	$v_{\text{max}} = \frac{qBR}{m}$

Conceptual Questions

Exercise:

Problem:

Describe the primary function of the electric field and the magnetic field in a cyclotron.

Problems

Exercise:

Problem:

A physicist is designing a cyclotron to accelerate protons to one-tenth the speed of light. The magnetic field will have a strength of 1.5 T. Determine (a) the rotational period of the circulating protons and (b) the maximum radius of the protons’ orbit.

Solution:

a. $4.4 \times 10^{-8}\text{s}$; b. 0.21 m

Exercise:

Problem:

The strengths of the fields in the velocity selector of a Bainbridge mass spectrometer are B = 0.500 T and E = $1.2 \times 10^{5}\,\text{V/m}$, and the strength of the magnetic field that separates the ions is $B_o = 0.750\,\text{T}$. A stream of singly charged Li ions is found to bend in a circular arc of radius 2.32 cm. What is the mass of the Li ions?

Exercise:

Problem:

The magnetic field in a cyclotron is 1.25 T, and the maximum orbital radius of the circulating protons is 0.40 m. (a) What is the kinetic energy of the protons when they are ejected from the cyclotron? (b) What is this energy in MeV? (c) Through what potential difference would a proton have to be accelerated to acquire this kinetic energy? (d) What is the period of the voltage source used to accelerate the protons? (e) Repeat the calculations for alpha-particles.

Solution:

a. $1.92 \times 10^{-12}\text{J}$; b. 12 MeV; c. 12 MV; d. $5.2 \times 10^{-8}\text{s}$; e. $1.92 \times 10^{-12}\text{J}$, 12 MeV, 12 V, $10.4 \times 10^{-8}\text{s}$

Exercise:

Problem:

A mass spectrometer is being used to separate common oxygen-16 from the much rarer oxygen-18, taken from a sample of old glacial ice. (The relative abundance of these oxygen isotopes is related to climatic temperature at the time the ice was deposited.) The ratio of the masses of these two ions is 16 to 18, the mass of oxygen-16 is $2.66 \times 10^{-26}\text{kg}$, and they are singly charged and travel at $5.00 \times 10^{6}\text{m/s}$ in a 1.20-T magnetic field. What is the separation between their paths when they hit a target after traversing a semicircle?

Exercise:

Problem:

(a) Triply charged uranium-235 and uranium-238 ions are being separated in a mass spectrometer. (The much rarer uranium-235 is used as reactor fuel.) The masses of the ions are $3.90 \times 10^{-25}\mathrm{kg}$ and $3.95 \times 10^{-25}\mathrm{kg}$, respectively, and they travel at $3.0 \times 10^{5}\mathrm{m/s}$ in a 0.250-T field. What is the separation between their paths when they hit a target after traversing a semicircle? (b) Discuss whether this distance between their paths seems to be big enough to be practical in the separation of uranium-235 from uranium-238.

Solution:

a. $2.50 \times 10^{-2}\mathrm{m}$; b. Yes, this distance between their paths is clearly big enough to separate the U-235 from the U-238, since it is a distance of 2.5 cm.

Additional Problems

Exercise:

Problem:

Calculate the magnetic force on a hypothetical particle of charge $1.0 \times 10^{-19}\mathrm{C}$ moving with a velocity of $6.0 \times 10^{4}\hat{\mathbf{i}}\mathrm{m/s}$ in a magnetic field of $1.2\widehat{\mathbf{k}}\mathrm{T}$.

Exercise:

Problem: Repeat the previous problem with a new magnetic field of $(0.4\hat{\mathbf{i}} + 1.2\widehat{\mathbf{k}})\mathrm{T}$.

Solution:

$-7.2 \times 10^{-15}\mathrm{N}\hat{\mathbf{j}}$

Exercise:

Problem:

An electron is projected into a uniform magnetic field $(0.5\hat{\mathbf{i}} + 0.8\widehat{\mathbf{k}})\mathrm{T}$ with a velocity of $(3.0\hat{\mathbf{i}} + 4.0\hat{\mathbf{j}}) \times 10^{6}\ \mathrm{m/s}$. What is the magnetic force on the electron?

Exercise:

Problem:

The mass and charge of a water droplet are $1.0 \times 10^{-4}\mathrm{g}$ and $2.0 \times 10^{-8}\mathrm{C}$, respectively. If the droplet is given an initial horizontal velocity of $5.0 \times 10^{5}\hat{\mathbf{i}}\mathrm{m/s}$, what magnetic field will keep it moving in this direction? Why must gravity be considered here?

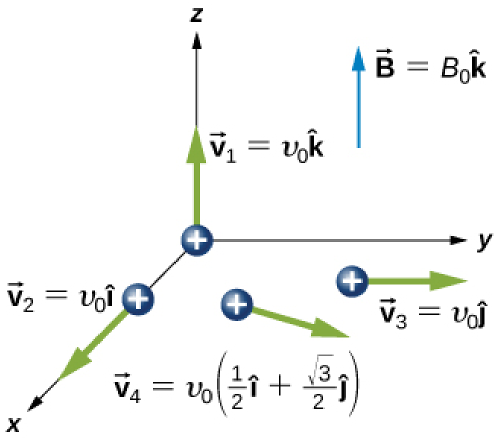
Solution:

$9.8 \times 10^{-5}\hat{\mathbf{j}}\mathrm{T}$; the magnetic and gravitational forces must balance to maintain dynamic equilibrium

Exercise:

Problem:

Four different proton velocities are given. For each case, determine the magnetic force on the proton in terms of e, v_0, and B_0.


z
$\vec{\mathbf{B}} = B_0\hat{\mathbf{k}}$
$\vec{\mathbf{v}}_1 = v_0\hat{\mathbf{k}}$
y
$\vec{\mathbf{v}}_2 = v_0\hat{\mathbf{i}}$
$\vec{\mathbf{v}}_3 = v_0\hat{\mathbf{j}}$
$\vec{\mathbf{v}}_4 = v_0\left(\frac{1}{2}\hat{\mathbf{i}} + \frac{\sqrt{3}}{2}\hat{\mathbf{j}}\right)$
x

Exercise:

Problem:

An electron of kinetic energy 2000 eV passes between parallel plates that are 1.0 cm apart and kept at a potential difference of 300 V. What is the strength of the uniform magnetic field B that will allow the electron to travel undeflected through the plates? Assume E and B are perpendicular.

Solution:

$1.13 \times 10^{-3}\text{T}$

Exercise:

Problem:

An alpha-particle $\left(m = 6.64 \times 10^{-27}\text{kg},\ q = 3.2 \times 10^{-19}\text{C}\right)$ moving with a velocity $\vec{\mathbf{v}} = (2.0\hat{\mathbf{i}} - 4.0\widehat{\mathbf{k}}) \times 10^{6}\text{m/s}$ enters a region where $\vec{\mathbf{E}} = (5.0\hat{\mathbf{i}} - 2.0\hat{\mathbf{j}}) \times 10^{4}\ \text{V/m}$ and $\vec{\mathbf{B}} = (1.0\hat{\mathbf{i}} + 4.0\widehat{\mathbf{k}}) \times 10^{-2}\text{T}$. What is the initial force on it?

Exercise:

Problem:

An electron moving with a velocity $\vec{\mathbf{v}} = \left(4.0\hat{\mathbf{i}} + 3.0\hat{\mathbf{j}} + 2.0\widehat{\mathbf{k}}\right) \times 10^{6}\text{m/s}$ enters a region where there is a uniform electric field and a uniform magnetic field. The magnetic field is given by $\vec{B} = \left(1.0\hat{\mathbf{i}} - 2.0\hat{\mathbf{j}} + 4.0\widehat{\mathbf{k}}\right) \times 10^{-2}\text{T}$. If the electron travels through a region without being deflected, what is the electric field?

Solution:

$(1.6\hat{\mathbf{i}} - 1.4\hat{\mathbf{j}} - 1.1\widehat{\mathbf{k}}) \times 10^5 \text{ V/m}$

Exercise:

Problem:

At a particular instant, an electron is traveling west to east with a kinetic energy of 10 keV. Earth's magnetic field has a horizontal component of $1.8 \times 10^{-5}\text{T}$ north and a vertical component of $5.0 \times 10^{-5}\text{T}$ down. (a) What is the path of the electron? (b) What is the radius of curvature of the path?

Exercise:

Problem:

What is the (a) path of a proton and (b) the magnetic force on the proton that is traveling west to east with a kinetic energy of 10 keV in Earth's magnetic field that has a horizontal component of 1.8 x 10^{-5} T north and a vertical component of 5.0 x 10^{-5} T down?

Solution:

a. circular motion in a north, down plane; b. $(1.61\hat{\mathbf{j}} - 0.58\widehat{\mathbf{k}}) \times 10^{-14}\text{N}$

Exercise:

Problem:

What magnetic field is required in order to confine a proton moving with a speed of $4.0 \times 10^6\text{m/s}$ to a circular orbit of radius 10 cm?

Exercise:

Problem:

An electron and a proton move with the same speed in a plane perpendicular to a uniform magnetic field. Compare the radii and periods of their orbits.

Solution:

The proton has more mass than the electron; therefore, its radius and period will be larger.

Exercise:

Problem:

A proton and an alpha-particle have the same kinetic energy and both move in a plane perpendicular to a uniform magnetic field. Compare the periods of their orbits.

Exercise:

Problem:

A singly charged ion takes $2.0 \times 10^{-3}\text{s}$ to complete eight revolutions in a uniform magnetic field of magnitude $2.0 \times 10^{-2}\text{T}$. What is the mass of the ion?

Solution:

$1.3 \times 10^{-25}\mathrm{kg}$

Exercise:

Problem:

A particle moving downward at a speed of $6.0 \times 10^{6}\mathrm{m/s}$ enters a uniform magnetic field that is horizontal and directed from east to west. (a) If the particle is deflected initially to the north in a circular arc, is its charge positive or negative? (b) If *B* = 0.25 T and the charge-to-mass ratio (*q/m*) of the particle is $4.0 \times 10^{7}\mathrm{C/kg}$, what is the radius of the path? (c) What is the speed of the particle after it has moved in the field for $1.0 \times 10^{-5}\mathrm{s}$? for 2.0 s?

Exercise:

Problem:

A proton, deuteron, and an alpha-particle are all accelerated from rest through the same potential difference. They then enter the same magnetic field, moving perpendicular to it. Compute the ratios of the radii of their circular paths. Assume that $m_d = 2m_p$ and $m_\alpha = 4m_p$.

Solution:

1:0.707:1

Exercise:

Problem:

A singly charged ion is moving in a uniform magnetic field of $7.5 \times 10^{-2}\mathrm{T}$ completes 10 revolutions in $3.47 \times 10^{-4}\mathrm{s}$. Identify the ion.

Exercise:

Problem:

Two particles have the same linear momentum, but particle A has four times the charge of particle B. If both particles move in a plane perpendicular to a uniform magnetic field, what is the ratio R_A/R_B of the radii of their circular orbits?

Solution:

1/4

Exercise:

Problem:

A uniform magnetic field of magnitude B is directed parallel to the *z*-axis. A proton enters the field with a velocity $\vec{\mathbf{v}} = (4\hat{\mathbf{j}} + 3\widehat{\mathbf{k}}) \times 10^{6}\mathrm{m/s}$ and travels in a helical path with a radius of 5.0 cm. (a) What is the value of B? (b) What is the time required for one trip around the helix? (c) Where is the proton $5.0 \times 10^{-7}\mathrm{s}$ after entering the field?

Exercise:

Problem:

An electron moving along the +*x* -axis at $5.0\ \times\ 10^{6}\text{m/s}$ enters a magnetic field that makes a 75^{o} angle with the *x*-axis of magnitude 0.20 T. Calculate the (a) pitch and (b) radius of the trajectory.

Solution:

a. $2.3\ \times\ 10^{-4}\text{m}$; b. $1.37\ \times\ 10^{-4}\text{m}$

Exercise:

Problem:

(a) A 0.750-m-long section of cable carrying current to a car starter motor makes an angle of 60º with Earth's $5.5\ \times\ 10^{-5}\text{T}$ field. What is the current when the wire experiences a force of $7.0\ \times\ 10^{-3}\text{N}$? (b) If you run the wire between the poles of a strong horseshoe magnet, subjecting 5.00 cm of it to a 1.75-T field, what force is exerted on this segment of wire?

Exercise:

Problem:

(a) What is the angle between a wire carrying an 8.00-A current and the 1.20-T field it is in if 50.0 cm of the wire experiences a magnetic force of 2.40 N? (b) What is the force on the wire if it is rotated to make an angle of 90º with the field?

Solution:

a. 30.0°; b. 4.80 N

Exercise:

Problem:

A 1.0-m-long segment of wire lies along the *x*-axis and carries a current of 2.0 A in the positive *x*-direction. Around the wire is the magnetic field of $\left(3.0\hat{\mathbf{i}}\ \times\ 4.0\widehat{\mathbf{k}}\right)\ \times\ 10^{-3}\text{T}$. Find the magnetic force on this segment.

Exercise:

Problem:

A 5.0-m section of a long, straight wire carries a current of 10 A while in a uniform magnetic field of magnitude $8.0\ \times\ 10^{-3}\text{T}$. Calculate the magnitude of the force on the section if the angle between the field and the direction of the current is (a) 45°; (b) 90°; (c) 0°; or (d) 180°.

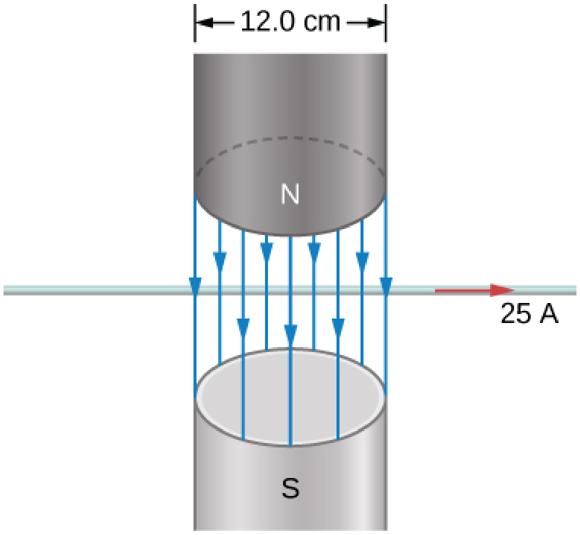
Solution:

a. 0.283 N; b. 0.4 N; c. 0 N; d. 0 N

Exercise:

Problem:

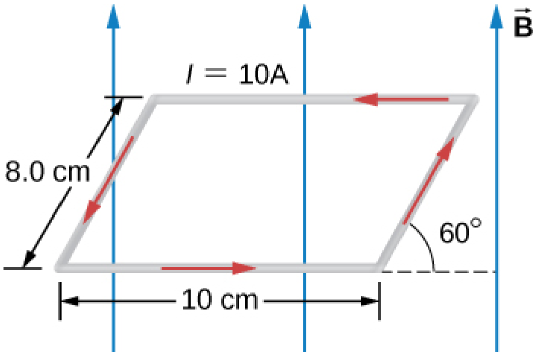
An electromagnet produces a magnetic field of magnitude 1.5 T throughout a cylindrical region of radius 6.0 cm. A straight wire carrying a current of 25 A passes through the field as shown in the accompanying figure. What is the magnetic force on the wire?



Exercise:

Problem:

The current loop shown in the accompanying figure lies in the plane of the page, as does the magnetic field. Determine the net force and the net torque on the loop if *I* = 10 A and *B* = 1.5 T.



Solution:

0 N and 0.012 Nm

Exercise:

Problem:

A circular coil of radius 5.0 cm is wound with five turns and carries a current of 5.0 A. If the coil is placed in a uniform magnetic field of strength 5.0 T, what is the maximum torque on it?

Exercise:

Problem:

A circular coil of wire of radius 5.0 cm has 20 turns and carries a current of 2.0 A. The coil lies in a magnetic field of magnitude 0.50 T that is directed parallel to the plane of the coil. (a) What is the magnetic dipole moment of the coil? (b) What is the torque on the coil?

Solution:

a. $0.31\ \mathrm{Am}^2$; b. 0.16 Nm

Exercise:

Problem:

A current-carrying coil in a magnetic field experiences a torque that is 75% of the maximum possible torque. What is the angle between the magnetic field and the normal to the plane of the coil?

Exercise:

Problem:

A 4.0-cm by 6.0-cm rectangular current loop carries a current of 10 A. What is the magnetic dipole moment of the loop?

Solution:

$0.024\ \mathrm{Am}^2$

Exercise:

Problem:

A circular coil with 200 turns has a radius of 2.0 cm. (a) What current through the coil results in a magnetic dipole moment of 3.0 Am^2? (b) What is the maximum torque that the coil will experience in a uniform field of strength $5.0 \times 10^{-2}\mathrm{T}$? (c) If the angle between μ and B is 45°, what is the magnitude of the torque on the coil? (d) What is the magnetic potential energy of coil for this orientation?

Exercise:

Problem:

The current through a circular wire loop of radius 10 cm is 5.0 A. (a) Calculate the magnetic dipole moment of the loop. (b) What is the torque on the loop if it is in a uniform 0.20-T magnetic field such that μ and B are directed at 30° to each other? (c) For this position, what is the potential energy of the dipole?

Solution:

a. $0.16\ \mathrm{Am}^2$; b. 0.016 Nm; c. 0.028 J

Exercise:

Problem:

A wire of length 1.0 m is wound into a single-turn planar loop. The loop carries a current of 5.0 A, and it is placed in a uniform magnetic field of strength 0.25 T. (a) What is the maximum torque that the loop will experience if it is square? (b) If it is circular? (c) At what angle relative to *B* would the normal to the circular coil have to be oriented so that the torque on it would be the same as the maximum torque on the square coil?

Exercise:

Problem:

Consider an electron rotating in a circular orbit of radius r. Show that the magnitudes of the magnetic dipole moment μ and the angular momentum *L* of the electron are related by:

Equation:

$$\frac{\mu}{L} = \frac{e}{2m}.$$

Solution:

(Proof)

Exercise:

Problem:

The Hall effect is to be used to find the sign of charge carriers in a semiconductor sample. The probe is placed between the poles of a magnet so that magnetic field is pointed up. A current is passed through a rectangular sample placed horizontally. As current is passed through the sample in the east direction, the north side of the sample is found to be at a higher potential than the south side. Decide if the number density of charge carriers is positively or negatively charged.

Exercise:

Problem:

The density of charge carriers for copper is 8.47×10^{28} electrons per cubic meter. What will be the Hall voltage reading from a probe made up of $3\,\text{cm} \times 2\,\text{cm} \times 1\,\text{cm}\ (\text{L} \times \text{W} \times \text{T})$ copper plate when a current of 1.5 A is passed through it in a magnetic field of 2.5 T perpendicular to the $3\,\text{cm} \times 2\,\text{cm}$.

Solution:

$4.65 \times 10^{-7}\text{V}$

Exercise:

Problem:

The Hall effect is to be used to find the density of charge carriers in an unknown material. A Hall voltage 40 μV for 3-A current is observed in a 3-T magnetic field for a rectangular sample with length 2 cm, width 1.5 cm, and height 0.4 cm. Determine the density of the charge carriers.

Exercise:

Problem:

Show that the Hall voltage across wires made of the same material, carrying identical currents, and subjected to the same magnetic field is inversely proportional to their diameters. (Hint: Consider how drift velocity depends on wire diameter.)

Solution:

Since $E = Blv$, where the width is twice the radius, $I = 2r$, $I = nqAv_d$,
$v_{\mathrm{d}} = \frac{I}{nqA} = \frac{I}{nq\pi r^2}$ so $E = B \times 2r \times \frac{\mathrm{I}}{nq\pi r^2} = \frac{2IB}{nq\pi r} \propto \frac{1}{r} \propto \frac{1}{d}$.
The Hall voltage is inversely proportional to the diameter of the wire.

Exercise:

Problem:

A velocity selector in a mass spectrometer uses a 0.100-T magnetic field. (a) What electric field strength is needed to select a speed of $4.0 \times 10^6 \mathrm{m/s}$? (b) What is the voltage between the plates if they are separated by 1.00 cm?

Exercise:

Problem:

Find the radius of curvature of the path of a 25.0-MeV proton moving perpendicularly to the 1.20-T field of a cyclotron.

Solution:

$6.92 \times 10^7 \mathrm{~m/s}$; 0.602 m

Exercise:

Problem:

Unreasonable results To construct a non-mechanical water meter, a 0.500-T magnetic field is placed across the supply water pipe to a home and the Hall voltage is recorded. (a) Find the flow rate through a 3.00-cm-diameter pipe if the Hall voltage is 60.0 mV. (b) What would the Hall voltage be for the same flow rate through a 10.0-cm-diameter pipe with the same field applied?

Exercise:

Problem:

Unreasonable results A charged particle having mass $6.64 \times 10^{-27}\text{kg}$ (that of a helium atom) moving at $8.70 \times 10^{5}\text{m/s}$ perpendicular to a 1.50-T magnetic field travels in a circular path of radius 16.0 mm. (a) What is the charge of the particle? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Solution:

a. $2.4 \times 10^{-19}\text{C}$; b. not an integer multiple of e; c. need to assume all charges have multiples of e, could be other forces not accounted for

Exercise:

Problem:

Unreasonable results An inventor wants to generate 120-V power by moving a 1.00-m-long wire perpendicular to Earth's $5.00 \times 10^{-5}\text{ T}$ field. (a) Find the speed with which the wire must move. (b) What is unreasonable about this result? (c) Which assumption is responsible?

Exercise:

Problem:

Unreasonable results Frustrated by the small Hall voltage obtained in blood flow measurements, a medical physicist decides to increase the applied magnetic field strength to get a 0.500-V output for blood moving at 30.0 cm/s in a 1.50-cm-diameter vessel. (a) What magnetic field strength is needed? (b) What is unreasonable about this result? (c) Which premise is responsible?

Solution:

a. B = 5 T; b. very large magnet; c. applying such a large voltage

Challenge Problems

Exercise:

Problem:

A particle of charge +*q* and mass *m* moves with velocity $\vec{\mathbf{v}}_0$ pointed in the +*y*-direction as it crosses the *x*-axis at *x* = *R* at a particular time. There is a negative charge –*Q* fixed at the origin, and there exists a uniform magnetic field $\vec{\mathbf{B}}_0$ pointed in the +*z*-direction. It is found that the particle describes a circle of radius *R* about –*Q*. Find $\vec{\mathbf{B}}_0$ in terms of the given quantities.

Exercise:

Problem:

A proton of speed $v = 6 \times 10^{5}\text{m/s}$ enters a region of uniform magnetic field of *B* = 0.5 T at an angle of $q = 30^{\circ}$ to the magnetic field. In the region of magnetic field proton describes a helical path with radius *R* and pitch *p* (distance between loops). Find *R* and *p*.

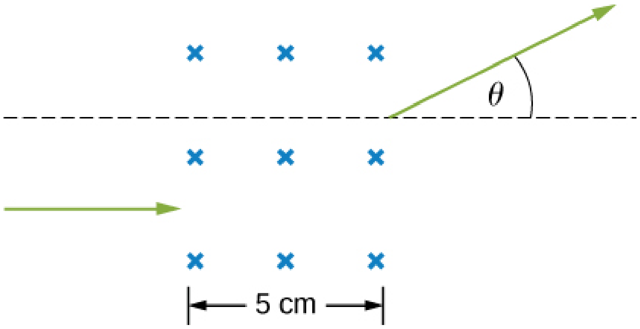
Solution:

$R = (mv\sin\theta)/qB; p = \left(\frac{2\pi m}{eB}\right)v\cos\theta$

Exercise:

Problem:

A particle’s path is bent when it passes through a region of non-zero magnetic field although its speed remains unchanged. This is very useful for “beam steering” in particle accelerators. Consider a proton of speed $4 \times 10^6 \text{m/s}$ entering a region of uniform magnetic field 0.2 T over a 5-cm-wide region. Magnetic field is perpendicular to the velocity of the particle. By how much angle will the path of the proton be bent? (Hint: The particle comes out tangent to a circle.)



Exercise:

Problem:

In a region a non-uniform magnetic field exists such that $B_x = 0, B_y = 0, \ and \ B_z = ax$, where *a* is a constant. At some time *t*, a wire of length *L* is carrying a current *I* is located along the *x*-axis from origin to *x* = *L*. Find the magnetic force on the wire at this instant in time.

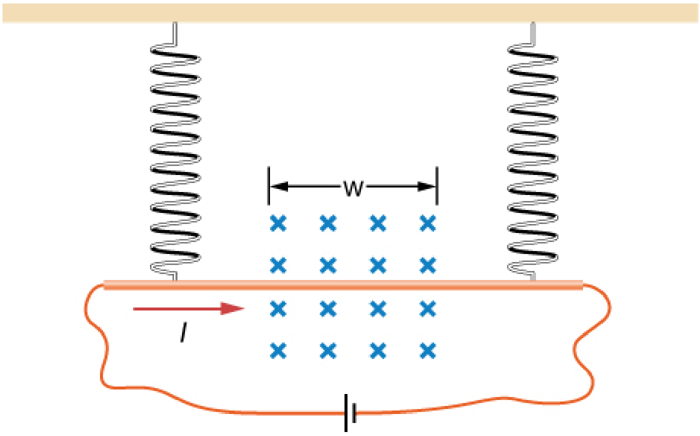
Solution:

$IaL^2/2$

Exercise:

Problem:

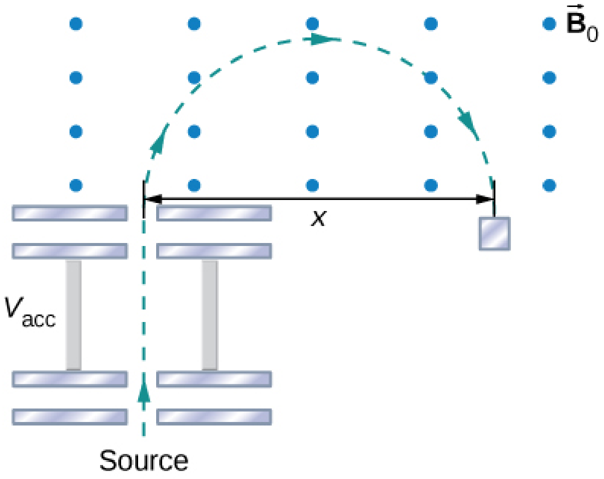
A copper rod of mass *m* and length *L* is hung from the ceiling using two springs of spring constant *k*. A uniform magnetic field of magnitude B_0 pointing perpendicular to the rod and spring (coming out of the page in the figure) exists in a region of space covering a length *w* of the copper rod. The ends of the rod are then connected by flexible copper wire across the terminals of a battery of voltage *V*. Determine the change in the length of the springs when *a* current *I* runs through the copper rod in the direction shown in figure. (Ignore any force by the flexible wire.)



Exercise:

Problem:

The accompanied figure shows an arrangement for measuring mass of ions by an instrument called the mass spectrometer. An ion of mass *m* and charge +*q* is produced essentially at rest in source *S*, a chamber in which a gas discharge is taking place. The ion is accelerated by a potential difference V_{acc} and allowed to enter a region of constant magnetic field $\vec{\mathbf{B}}_0$. In the uniform magnetic field region, the ion moves in a semicircular path striking a photographic plate at a distance *x* from the entry point. Derive a formula for mass m in terms of B_0, q, V_{acc}, and *x*.



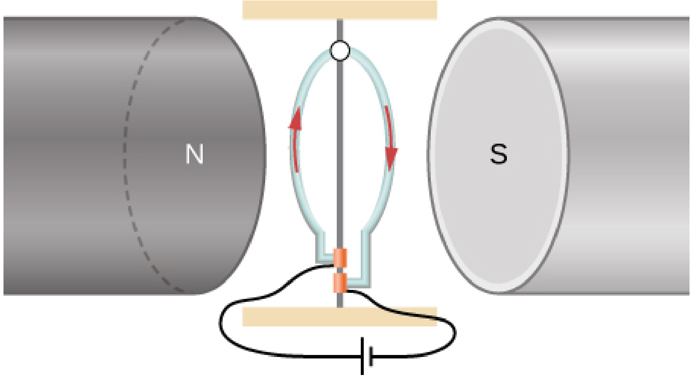
Solution:

$$m = \frac{qB_0{}^2}{8V_{\text{acc}}}x^2$$

Exercise:

Problem:

A wire is made into a circular shape of radius R and pivoted along a central support. The two ends of the wire are touching a brush that is connected to a dc power source. The structure is between the poles of a magnet such that we can assume there is a uniform magnetic field on the wire. In terms of a coordinate system with origin at the center of the ring, magnetic field is $B_x = B_0, B_y = B_z = 0$, and the ring rotates about the *z*-axis. Find the torque on the ring when it is not in the *xz*-plane.



Exercise:

Problem:

A long-rigid wire lies along the *x*-axis and carries a current of 2.5 A in the positive *x*-direction. Around the wire is the magnetic field $\vec{\mathbf{B}} = 2.0\hat{\mathbf{i}} + 5.0x^2\hat{\mathbf{j}}$, with *x* in meters and *B* in millitesla. Calculate the magnetic force on the segment of wire between *x* = 2.0 m and *x* = 4.0 m.

Solution:

0.23 N

Exercise:

Problem:

A circular loop of wire of area 10 cm^2 carries a current of 25 A. At a particular instant, the loop lies in the *xy*-plane and is subjected to a magnetic field $\vec{\mathbf{B}} = \left(2.0\hat{\mathbf{i}} + 6.0\hat{\mathbf{j}} + 8.0\widehat{\mathbf{k}}\right) \times 10^{-3}\,\text{T}$. As viewed from above the *xy*-plane, the current is circulating clockwise. (a) What is the magnetic dipole moment of the current loop? (b) At this instant, what is the magnetic torque on the loop?

Glossary

cyclotron
: device used to accelerate charged particles to large kinetic energies

dees

large metal containers used in cyclotrons that serve contain a stream of charged particles as their speed is increased

mass spectrometer
device that separates ions according to their charge-to-mass ratios

Introduction class="introduction"

An external hard drive attached to a computer works by magnetically encoding information that can be stored or retrieved quickly. A key idea in the development of digital devices is the ability to produce and use magnetic fields in this way. (credit: modification of work by “Miss Karen”/Flickr)

In the preceding chapter, we saw that a moving charged particle produces a magnetic field. This connection between electricity and magnetism is exploited in electromagnetic devices, such as a computer hard drive. In fact, it is the underlying principle behind most of the technology in modern society, including telephones, television, computers, and the internet.

In this chapter, we examine how magnetic fields are created by arbitrary distributions of electric current, using the Biot-Savart law. Then we look at how current-carrying wires create magnetic fields and deduce the forces that arise between two current-carrying wires due to these magnetic fields. We also study the torques produced by the magnetic fields of current loops. We then generalize these results to an important law of electromagnetism, called Ampère’s law.

We examine some devices that produce magnetic fields from currents in geometries based on loops, known as solenoids and toroids. Finally, we look at how materials behave in magnetic fields and categorize materials based on their responses to magnetic fields.

The Biot-Savart Law

By the end of this section, you will be able to:

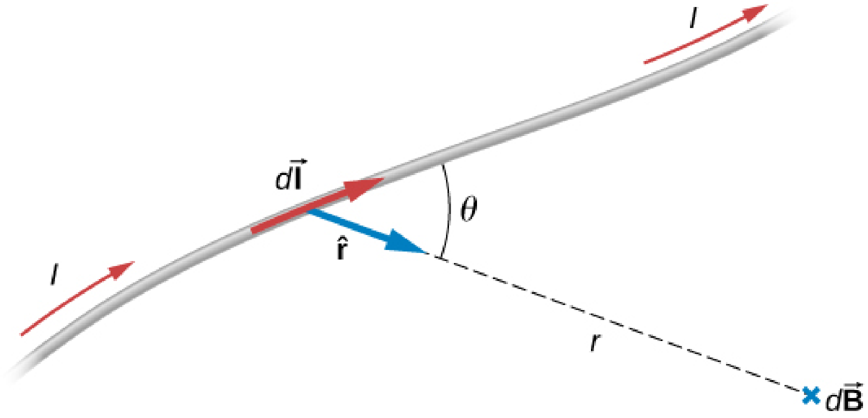
- Explain how to derive a magnetic field from an arbitrary current in a line segment
- Calculate magnetic field from the Biot-Savart law in specific geometries, such as a current in a line and a current in a circular arc

We have seen that mass produces a gravitational field and also interacts with that field. Charge produces an electric field and also interacts with that field. Since moving charge (that is, current) interacts with a magnetic field, we might expect that it also creates that field—and it does.

The equation used to calculate the magnetic field produced by a current is known as the Biot-Savart law. It is an empirical law named in honor of two scientists who investigated the interaction between a straight, current-carrying wire and a permanent magnet. This law enables us to calculate the magnitude and direction of the magnetic field produced by a current in a wire. The **Biot-Savart law** states that at any point *P* ([link]), the magnetic field $d\vec{\mathbf{B}}$ due to an element $d\vec{\mathbf{l}}$ of a current-carrying wire is given by

Equation:

$$d\vec{\mathbf{B}} = \frac{\mu_0}{4\pi}\frac{I d\vec{\mathbf{l}} \times \hat{\mathbf{r}}}{r^2}.$$



A current element $I d\vec{\mathbf{l}}$ produces a magnetic field at point *P* given by the Biot-Savart law.

The constant μ_0 is known as the **permeability of free space** and is exactly

Note:

Equation:

$$\mu_0 = 4\pi \times 10^{-7} \text{T} \cdot \text{m/A}$$

in the SI system. The infinitesimal wire segment $d\vec{\mathbf{l}}$ is in the same direction as the current *I* (assumed positive), *r* is the distance from $d\vec{\mathbf{l}}$ to *P* and $\hat{\mathbf{r}}$ is a unit vector that points from $d\vec{\mathbf{l}}$ to *P*, as shown in the figure.

The direction of $d\vec{\mathbf{B}}$ is determined by applying the right-hand rule to the vector product $d\vec{\mathbf{l}} \times \hat{\mathbf{r}}$. The magnitude of $d\vec{\mathbf{B}}$ is

Note:
Equation:

$$dB = \frac{\mu_0}{4\pi} \frac{I\, dl \, \sin\theta}{r^2}$$

where θ is the angle between $d\vec{\mathbf{l}}$ and $\hat{\mathbf{r}}$. Notice that if $\theta = 0$, then $d\vec{\mathbf{B}} = \vec{\mathbf{0}}$. The field produced by a current element $Id\vec{\mathbf{l}}$ has no component parallel to $d\vec{\mathbf{l}}$.

The magnetic field due to a finite length of current-carrying wire is found by integrating [link] along the wire, giving us the usual form of the Biot-Savart law.

Note:
Biot-Savart law

The magnetic field $\vec{\mathbf{B}}$ due to an element $d\vec{\mathbf{l}}$ of a current-carrying wire is given by
Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \int_{\text{wire}} \frac{I\, d\vec{\mathbf{l}} \times \hat{\mathbf{r}}}{r^2}.$$

Since this is a vector integral, contributions from different current elements may not point in the same direction. Consequently, the integral is often difficult to evaluate, even for fairly simple geometries. The following strategy may be helpful.

Note:

Problem-Solving Strategy: Solving Biot-Savart Problems

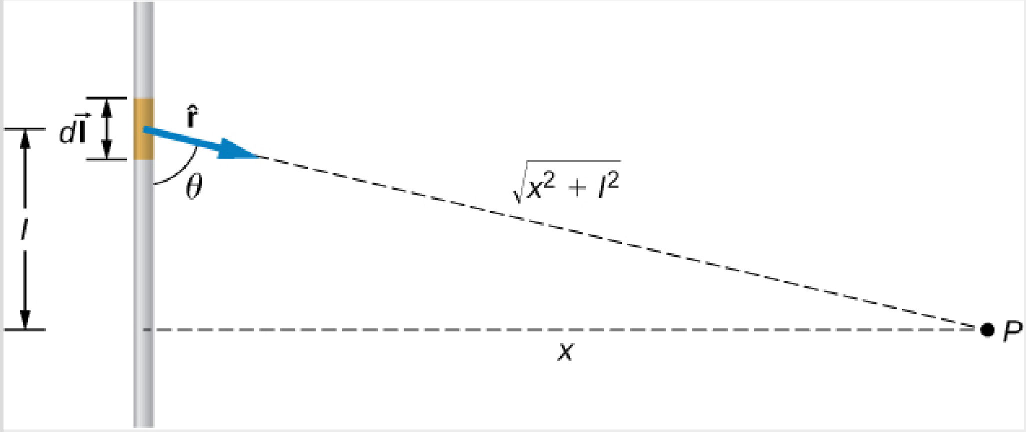
To solve Biot-Savart law problems, the following steps are helpful:

1. Identify that the Biot-Savart law is the chosen method to solve the given problem. If there is symmetry in the problem comparing $\vec{\mathbf{B}}$ and $\mathbf{d}\vec{\mathbf{l}}$, Ampère’s law may be the preferred method to solve the question.
2. Draw the current element length $d\vec{\mathbf{l}}$ and the unit vector $\hat{\mathbf{r}}$, noting that $d\vec{\mathbf{l}}$ points in the direction of the current and $\hat{\mathbf{r}}$ points from the current element toward the point where the field is desired.
3. Calculate the cross product $d\vec{\mathbf{l}} \times \hat{\mathbf{r}}$. The resultant vector gives the direction of the magnetic field according to the Biot-Savart law.
4. Use [link] and substitute all given quantities into the expression to solve for the magnetic field. Note all variables that remain constant over the entire length of the wire may be factored out of the integration.
5. Use the right-hand rule to verify the direction of the magnetic field produced from the current or to write down the direction of the magnetic field if only the magnitude was solved for in the previous part.

Example:

Calculating Magnetic Fields of Short Current Segments

A short wire of length 1.0 cm carries a current of 2.0 A in the vertical direction ([link]). The rest of the wire is shielded so it does not add to the magnetic field produced by the wire. Calculate the magnetic field at point *P*, which is 1 meter from the wire in the *x*-direction.



A small line segment carries a current *I* in the vertical direction. What is the magnetic field at a distance *x* from the segment?

Strategy

We can determine the magnetic field at point *P* using the Biot-Savart law. Since the current segment is much smaller than the distance *x*, we can drop the integral from the expression. The integration is converted back into a summation, but only for small *dl*, which we now write as Δl. Another way to think about it is that each of the radius values is nearly the same, no matter where the current element is on the line segment, if Δl is small compared to *x*. The angle θ is calculated using a tangent function. Using the numbers given, we can calculate the magnetic field at *P*.

Solution

The angle between $\Delta\vec{\mathbf{l}}$ and $\hat{\mathbf{r}}$ is calculated from trigonometry, knowing the distances *l* and *x* from the problem:

Equation:

$$\theta = \tan^{-1}\left(\frac{1\ \text{m}}{0.01\ \text{m}}\right) = 89.4^{\circ}.$$

The magnetic field at point *P* is calculated by the Biot-Savart law:

Equation:

$$B = \frac{\mu_0}{4\pi}\frac{I\Delta l \sin\theta}{r^2} = \left(1 \times 10^{-7}\text{T}\cdot\text{m/A}\right)\left(\frac{2\ \text{A}(0.01\ \text{m})\sin(89.4^{\circ})}{(1\ \text{m})^2}\right) = 2.0 \times 10^{-9}\text{T}.$$

From the right-hand rule and the Biot-Savart law, the field is directed into the page.

Significance

This approximation is only good if the length of the line segment is very small compared to the distance from the current element to the point. If not, the integral form of the Biot-Savart law must be used over the entire line segment to calculate the magnetic field.

Note:

Exercise:

Problem:

Check Your Understanding Using [link], at what distance would *P* have to be to measure a magnetic field half of the given answer?

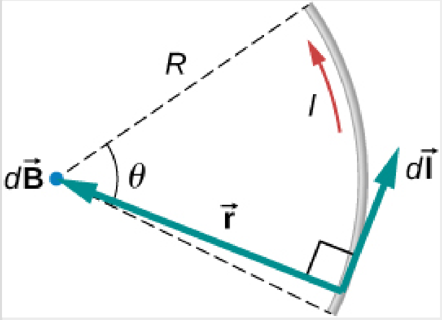
Solution:

1.41 meters

Example:

Calculating Magnetic Field of a Circular Arc of Wire

A wire carries a current I in a circular arc with radius R swept through an arbitrary angle θ ([link]). Calculate the magnetic field at the center of this arc at point P.



A wire segment carrying a current I. The path $d\vec{\mathbf{l}}$ and radial direction $\hat{\mathbf{r}}$ are indicated.

Strategy

We can determine the magnetic field at point P using the Biot-Savart law. The radial and path length directions are always at a right angle, so the cross product turns into multiplication. We also know that the distance along the path dl is related to the radius times the angle θ (in radians). Then we can pull all constants out of the integration and solve for the magnetic field.

Solution

The Biot-Savart law starts with the following equation:

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \int_{\text{wire}} \frac{I d\vec{\mathbf{l}} \times \hat{\mathbf{r}}}{r^2}.$$

As we integrate along the arc, all the contributions to the magnetic field are in the same direction (out of the page), so we can work with the magnitude of the field. The cross product turns into multiplication because the path dl and the radial direction are perpendicular. We can also substitute the arc length formula, $dl = rd\theta$:

Equation:

$$B = \frac{\mu_0}{4\pi} \int_{\text{wire}} \frac{Ir\, d\theta}{r^2}.$$

The current and radius can be pulled out of the integral because they are the same regardless of where we are on the path. This leaves only the integral over the angle,

Equation:

$$B = \frac{\mu_0 I}{4\pi r} \int_{\text{wire}} d\theta.$$

The angle varies on the wire from 0 to θ; hence, the result is

Equation:

$$B = \frac{\mu_0 I\theta}{4\pi r}.$$

Significance

The direction of the magnetic field at point *P* is determined by the right-hand rule, as shown in the previous chapter. If there are other wires in the diagram along with the arc, and you are asked to find the net magnetic field, find each contribution from a wire or arc and add the results by superposition of vectors. Make sure to pay attention to the direction of each contribution. Also note that in a symmetric situation, like a straight or circular wire, contributions from opposite sides of point *P* cancel each other.

Note:

Exercise:

Problem:

Check Your Understanding The wire loop forms a full circle of radius *R* and current *I*. What is the magnitude of the magnetic field at the center?

Solution:

$\frac{\mu_0 I}{2R}$

Summary

- The magnetic field created by a current-carrying wire is found by the Biot-Savart law.
- The current element $Id\vec{\mathbf{l}}$ produces a magnetic field a distance *r* away.

Conceptual Questions

Exercise:

Problem:

For calculating magnetic fields, what are the advantages and disadvantages of the Biot-Savart law?

Solution:

Biot-Savart law’s advantage is that it works with any magnetic field produced by a current loop. The disadvantage is that it can take a long time.

Exercise:

Problem:

Describe the magnetic field due to the current in two wires connected to the two terminals of a source of emf and twisted tightly around each other.

Exercise:

Problem: How can you decide if a wire is infinite?

Solution:

If you were to go to the start of a line segment and calculate the angle θ to be approximately 0°, the wire can be considered infinite. This judgment is based also on the precision you need in the result.

Exercise:

Problem:

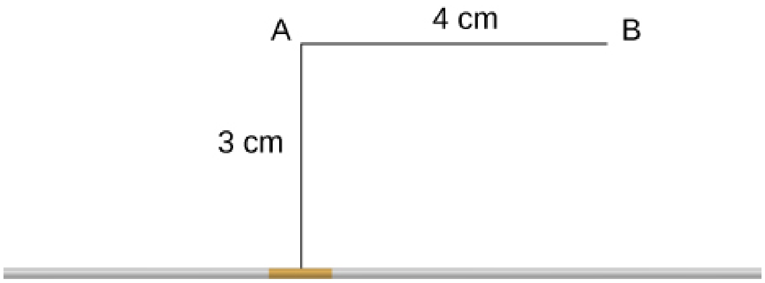
Identical currents are carried in two circular loops; however, one loop has twice the diameter as the other loop. Compare the magnetic fields created by the loops at the center of each loop.

Problems

Exercise:

Problem:

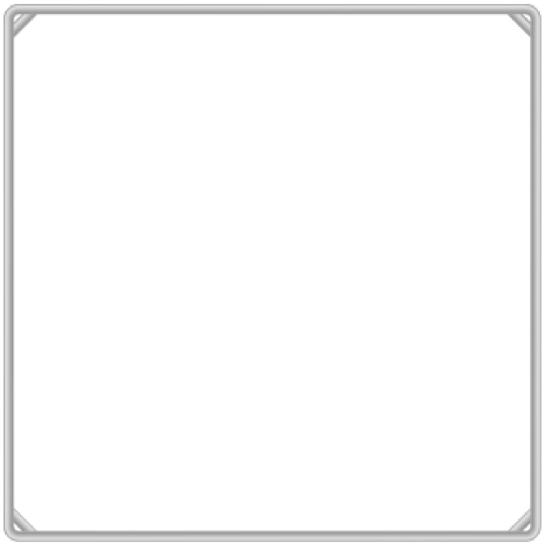
A 10-A current flows through the wire shown. What is the magnitude of the magnetic field due to a 0.5-mm segment of wire as measured at (a) point A and (b) point B?



Exercise:

Problem:

Ten amps flow through a square loop where each side is 20 cm in length. At each corner of the loop is a 0.01-cm segment that connects the longer wires as shown. Calculate the magnitude of the magnetic field at the center of the loop.

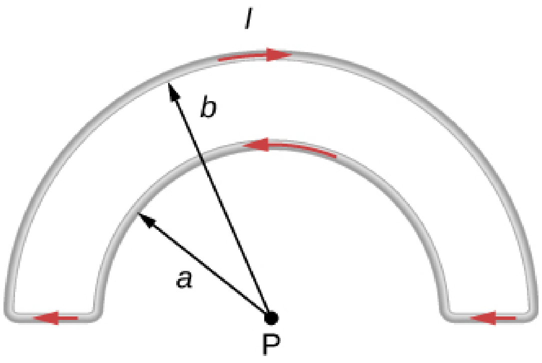


Solution:

$5.66 \times 10^{-5} \mathrm{T}$

Exercise:

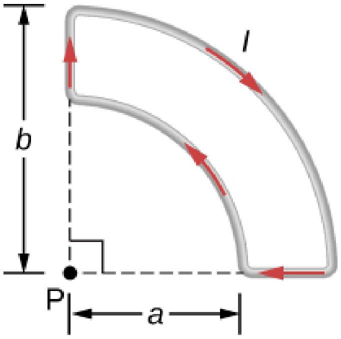
Problem: What is the magnetic field at P due to the current *I* in the wire shown?



Exercise:

Problem:

The accompanying figure shows a current loop consisting of two concentric circular arcs and two perpendicular radial lines. Determine the magnetic field at point P.



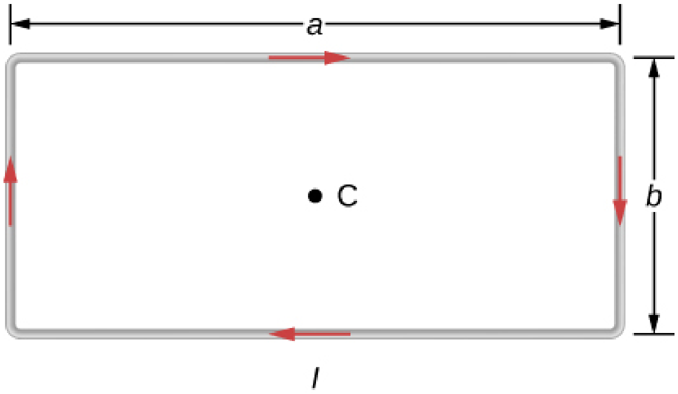
Solution:

$B = \frac{\mu_o I}{8} \ \frac{1}{a} - \frac{1}{b}$ out of the page

Exercise:

Problem:

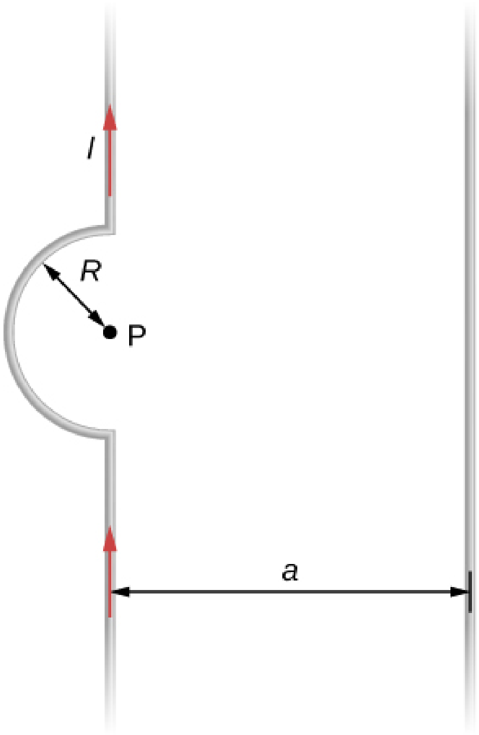
Find the magnetic field at the center C of the rectangular loop of wire shown in the accompanying figure.



Exercise:

Problem:

Two long wires, one of which has a semicircular bend of radius *R*, are positioned as shown in the accompanying figure. If both wires carry a current *I*, how far apart must their parallel sections be so that the net magnetic field at P is zero? Does the current in the straight wire flow up or down?



Solution:

$a = \frac{2R}{\pi}$; the current in the wire to the right must flow up the page.

Glossary

Biot-Savart law

an equation giving the magnetic field at a point produced by a current-carrying wire

permeability of free space

μ_0, measure of the ability of a material, in this case free space, to support a magnetic field

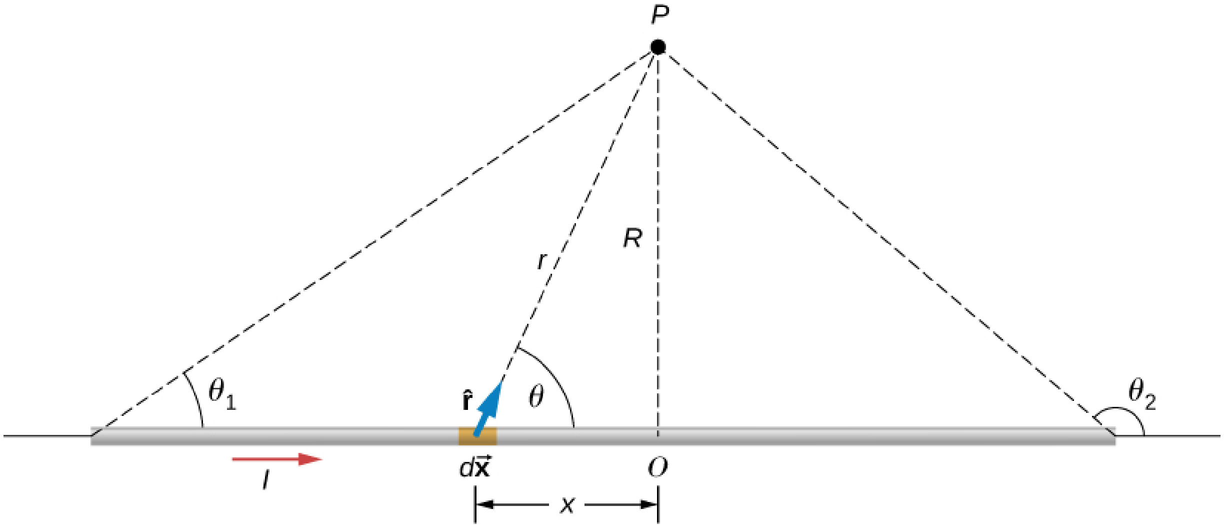
Magnetic Field Due to a Thin Straight Wire

By the end of this section, you will be able to:

- Explain how the Biot-Savart law is used to determine the magnetic field due to a thin, straight wire.
- Determine the dependence of the magnetic field from a thin, straight wire based on the distance from it and the current flowing in the wire.
- Sketch the magnetic field created from a thin, straight wire by using the second right-hand rule.

How much current is needed to produce a significant magnetic field, perhaps as strong as Earth’s field? Surveyors will tell you that overhead electric power lines create magnetic fields that interfere with their compass readings. Indeed, when Oersted discovered in 1820 that a current in a wire affected a compass needle, he was not dealing with extremely large currents. How does the shape of wires carrying current affect the shape of the magnetic field created? We noted in Chapter 28 that a current loop created a magnetic field similar to that of a bar magnet, but what about a straight wire? We can use the Biot-Savart law to answer all of these questions, including determining the magnetic field of a long straight wire.

[link] shows a section of an infinitely long, straight wire that carries a current *I*. What is the magnetic field at a point *P*, located a distance *R* from the wire?



A section of a thin, straight current-carrying wire. The independent variable θ has the limits θ_1 and θ_2.

Let's begin by considering the magnetic field due to the current element $I\,d\vec{\mathbf{x}}$ located at the position *x*. Using the right-hand rule 1 from the previous chapter, $d\vec{\mathbf{x}}\ \times\ \hat{\mathbf{r}}$ points out of the page for any element along the wire. At point *P*, therefore, the magnetic fields due to all current elements have the same direction. This means that we can calculate the net field there by evaluating the scalar sum of the contributions of the elements. With $\left|d\vec{\mathbf{x}}\ \times\ \hat{\mathbf{r}}\right| = (dx)(1)\sin\theta$, we have from the Biot-Savart law

Equation:

$$B = \frac{\mu_0}{4\pi}\int\limits_{\text{wire}}\frac{I\sin\theta\,dx}{r^2}.$$

The wire is symmetrical about point *O*, so we can set the limits of the integration from zero to infinity and double the answer, rather than integrate from negative infinity to positive infinity. Based on the picture and geometry, we can write expressions for *r* and sin θ in terms of *x* and *R*, namely:

Equation:

$$\begin{array}{rcl} r & = & \sqrt{x^2+R^2} \\ \sin\theta & = & \frac{R}{\sqrt{x^2+R^2}}. \end{array}$$

Substituting these expressions into [link], the magnetic field integration becomes

Equation:

$$B = \frac{\mu_o I}{2\pi}\int_0^\infty\frac{R\,dx}{\left(x^2+R^2\right)^{3/2}}.$$

Evaluating the integral yields

Equation:

$$B=\frac{\mu_o I}{2\pi R}\left[\frac{x}{\left(x^2+R^2\right)^{1/2}}\right]_0^\infty.$$

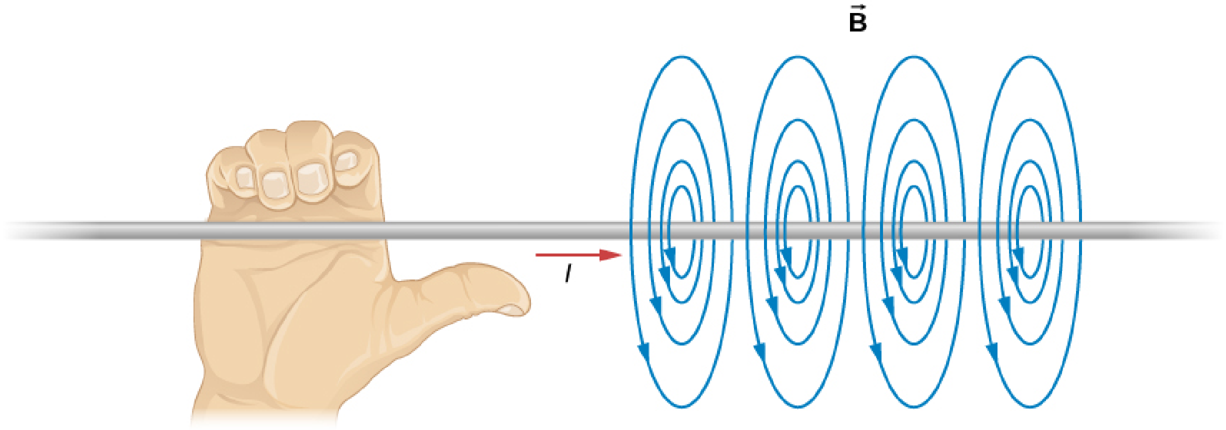
Substituting the limits gives us the solution

Note:

Equation:

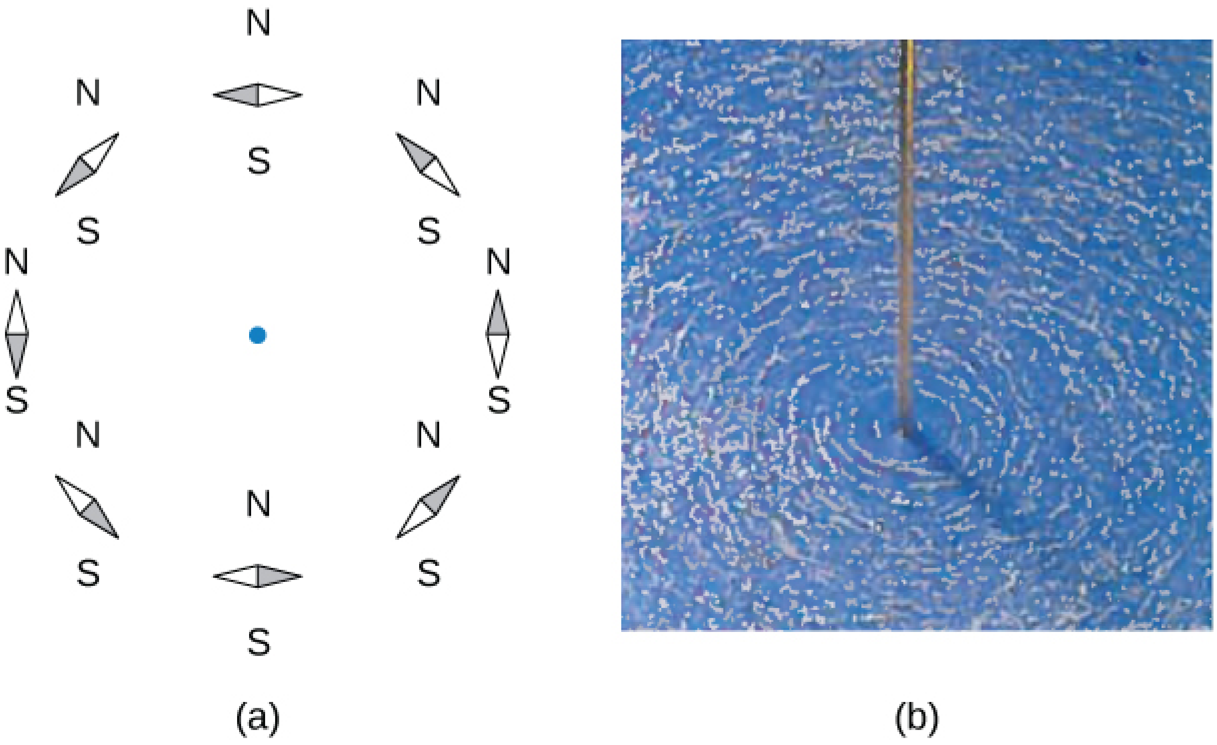
$$B=\frac{\mu_o I}{2\pi R}.$$

The magnetic field lines of the infinite wire are circular and centered at the wire ([link]), and they are identical in every plane perpendicular to the wire. Since the field decreases with distance from the wire, the spacing of the field lines must increase correspondingly with distance. The direction of this magnetic field may be found with a second form of the right-hand rule (illustrated in [link]). If you hold the wire with your right hand so that your thumb points along the current, then your fingers wrap around the wire in the same sense as $\vec{\mathbf{B}}$.



Some magnetic field lines of an infinite wire. The direction of $\vec{\mathbf{B}}$ can be found with a form of the right-hand rule.

The direction of the field lines can be observed experimentally by placing several small compass needles on a circle near the wire, as illustrated in [link]. When there is no current in the wire, the needles align with Earth’s magnetic field. However, when a large current is sent through the wire, the compass needles all point tangent to the circle. Iron filings sprinkled on a horizontal surface also delineate the field lines, as shown in [link].

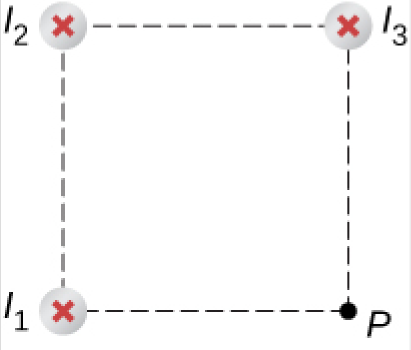


The shape of the magnetic field lines of a long wire can be seen using (a) small compass needles and (b) iron filings.

Example:

Calculating Magnetic Field Due to Three Wires

Three wires sit at the corners of a square, all carrying currents of 2 amps into the page as shown in [link]. Calculate the magnitude of the magnetic field at the other corner of the square, point *P*, if the length of each side of the square is 1 cm.



Three wires have current flowing into the page. The magnetic field is determined at the fourth corner of the square.

Strategy

The magnetic field due to each wire at the desired point is calculated. The diagonal distance is calculated using the Pythagorean theorem. Next, the direction of each magnetic field's contribution is determined by drawing a circle centered at the point of the wire and out toward the desired point. The direction of the magnetic field contribution from that wire is tangential to the curve. Lastly, working with these vectors, the resultant is calculated.

Solution

Wires 1 and 3 both have the same magnitude of magnetic field contribution at point *P*:

Equation:

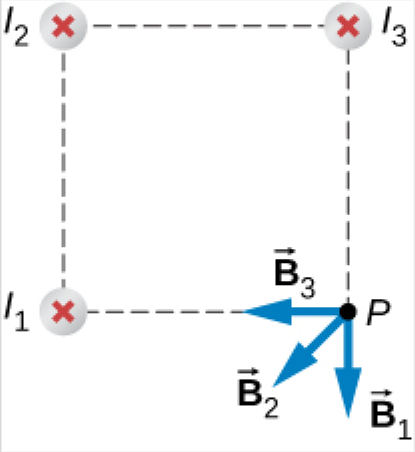
$$B_1 = B_3 = \frac{\mu_o I}{2\pi R} = \frac{(4\pi \times 10^{-7} \mathrm{T} \cdot \mathrm{m/A})(2\ \mathrm{A})}{2\pi(0.01\ \mathrm{m})} = 4 \times 10^{-5} \mathrm{T}.$$

Wire 2 has a longer distance and a magnetic field contribution at point *P* of:

Equation:

$$B_2 = \frac{\mu_o I}{2\pi R} = \frac{(4\pi \times 10^{-7} \text{T} \cdot \text{m/A})(2\,\text{A})}{2\pi(0.01414\,\text{m})} = 3 \times 10^{-5}\text{T}.$$

The vectors for each of these magnetic field contributions are shown.



The magnetic field in the *x*-direction has contributions from wire 3 and the *x*-component of wire 2:

Equation:

$$B_{\text{net } x} = -4 \times 10^{-5}\text{T} - 2.83 \times 10^{-5}\text{T} \cos(45^\circ) = -6 \times 10^{-5}\text{T}.$$

The *y*-component is similarly the contributions from wire 1 and the *y*-component of wire 2:

Equation:

$$B_{\text{net } y} = -4 \times 10^{-5}\text{T} - 2.83 \times 10^{-5}\text{T} \sin(45^\circ) = -6 \times 10^{-5}\text{T}.$$

Therefore, the net magnetic field is the resultant of these two components:

Equation:

$$\begin{aligned} B_{\text{net}} &= \sqrt{B_{\text{net } x}^2 + B_{\text{net } y}^2} \\ B_{\text{net}} &= \sqrt{(-6 \times 10^{-5}\text{T})^2 + (-6 \times 10^{-5}\text{T})^2} \\ B_{\text{net}} &= 8 \times 10^{-5}\text{T}. \end{aligned}$$

Significance
The geometry in this problem results in the magnetic field contributions in the *x*- and *y*-directions having the same magnitude. This is not necessarily the case if the currents were different values or if the wires were located in different positions. Regardless of the numerical results, working on the components of the vectors will yield the resulting magnetic field at the point in need.

Note:
Exercise:

Problem:

Check Your Understanding Using [link], keeping the currents the same in wires 1 and 3, what should the current be in wire 2 to counteract the magnetic fields from wires 1 and 3 so that there is no net magnetic field at point P?

Solution:

4 amps flowing out of the page

Summary

- The strength of the magnetic field created by current in a long straight wire is given by $B = \frac{\mu_0 I}{2\pi R}$ (long straight wire) where *I* is the current, *R* is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \ \times \ 10^{-7} \, \mathrm{T} \cdot \mathrm{m/s}$ is the permeability of free space.
- The direction of the magnetic field created by a long straight wire is given by right-hand rule 2 (RHR-2): Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops created by it.

Conceptual Questions

Exercise:

Problem:

How would you orient two long, straight, current-carrying wires so that there is no net magnetic force between them? (*Hint*: What orientation would lead to one wire not experiencing a magnetic field from the other?)

Solution:

You would make sure the currents flow perpendicular to one another.

Problems

Exercise:

Problem:

A typical current in a lightning bolt is 10^4 A. Estimate the magnetic field 1 m from the bolt.

Exercise:

Problem:

The magnitude of the magnetic field 50 cm from a long, thin, straight wire is $8.0\ \mu\text{T}$. What is the current through the long wire?

Solution:

20 A

Exercise:

Problem:

A transmission line strung 7.0 m above the ground carries a current of 500 A. What is the magnetic field on the ground directly below the wire? Compare your answer with the magnetic field of Earth.

Exercise:

Problem:

A long, straight, horizontal wire carries a left-to-right current of 20 A. If the wire is placed in a uniform magnetic field of magnitude $4.0 \times 10^{-5}\mathrm{T}$ that is directed vertically downward, what is the resultant magnitude of the magnetic field 20 cm above the wire? 20 cm below the wire?

Solution:

Both answers have the magnitude of magnetic field of $4.5 \times 10^{-5}\mathrm{T}$.

Exercise:

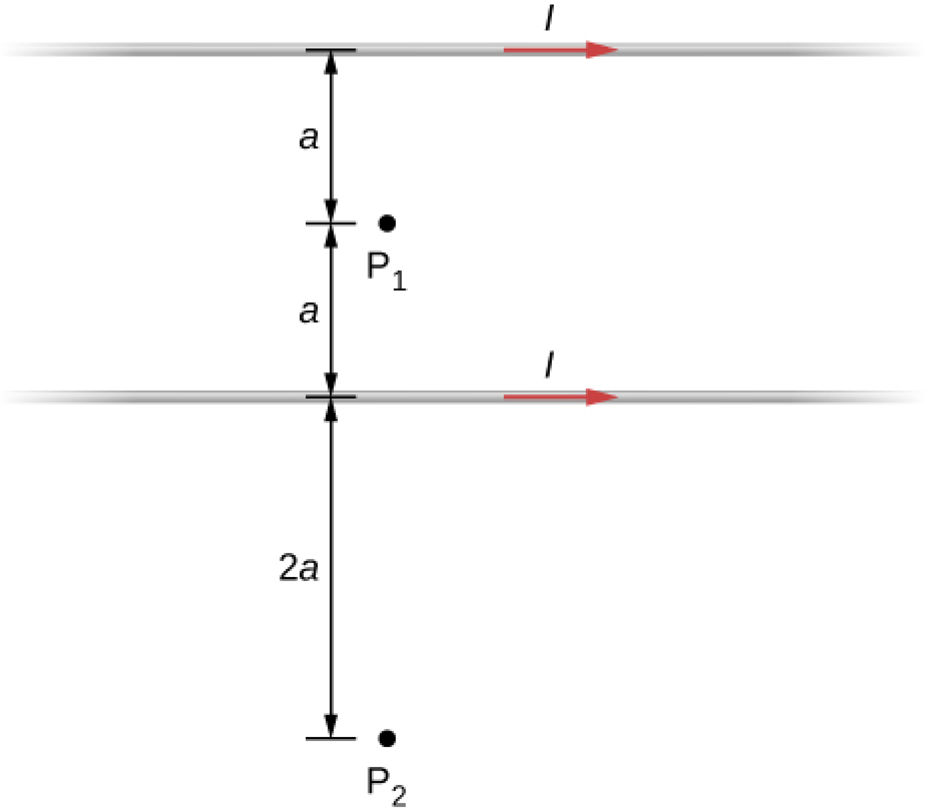
Problem:

The two long, parallel wires shown in the accompanying figure carry currents in the same direction. If $I_1 = 10\ \mathrm{A}$ and $I_2 = 20\ \mathrm{A}$, what is the magnetic field at point P?

Exercise:

Problem:

The accompanying figure shows two long, straight, horizontal wires that are parallel and a distance 2*a* apart. If both wires carry current *I* in the same direction, (a) what is the magnetic field at P_1? (b) P_2?



Solution:

At P1, the net magnetic field is zero. At P2, $B = \frac{3\mu_o I}{8\pi a}$ into the page.

Exercise:

Problem:

Repeat the calculations of the preceding problem with the direction of the current in the lower wire reversed.

Exercise:

Problem:

Consider the area between the wires of the preceding problem. At what distance from the top wire is the net magnetic field a minimum? Assume that the currents are equal and flow in opposite directions.

Solution:

The magnetic field is at a minimum at distance a from the top wire, or half-way between the wires.

Magnetic Force between Two Parallel Currents

By the end of this section, you will be able to:

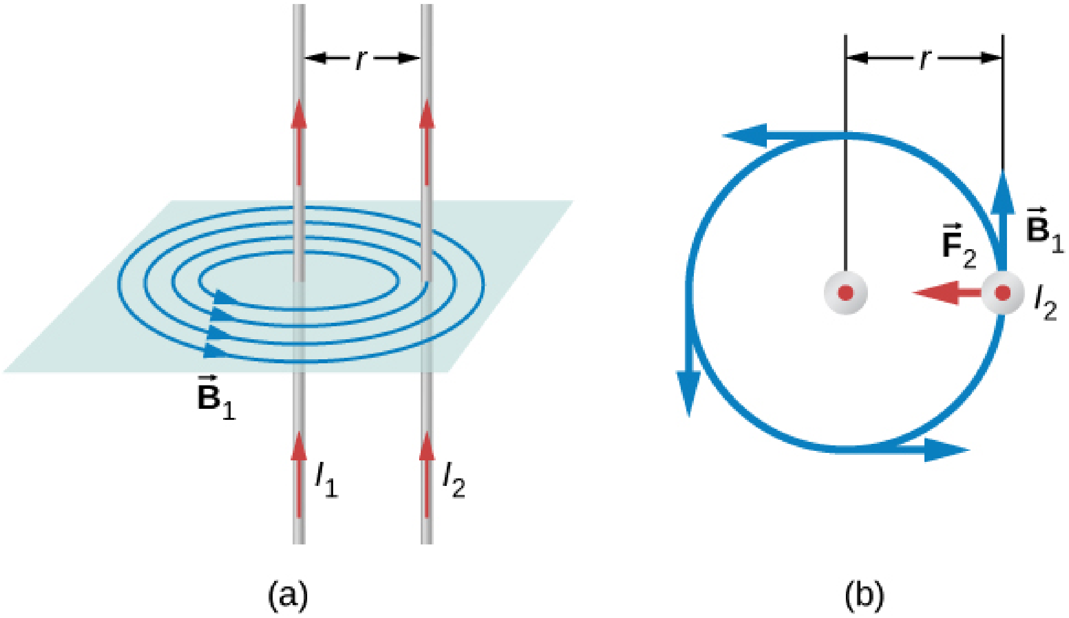
- Explain how parallel wires carrying currents can attract or repel each other
- Define the ampere and describe how it is related to current-carrying wires
- Calculate the force of attraction or repulsion between two current-carrying wires

You might expect that two current-carrying wires generate significant forces between them, since ordinary currents produce magnetic fields and these fields exert significant forces on ordinary currents. But you might not expect that the force between wires is used to define the ampere. It might also surprise you to learn that this force has something to do with why large circuit breakers burn up when they attempt to interrupt large currents.

The force between two long, straight, and parallel conductors separated by a distance *r* can be found by applying what we have developed in the preceding sections. [link] shows the wires, their currents, the field created by one wire, and the consequent force the other wire experiences from the created field. Let us consider the field produced by wire 1 and the force it exerts on wire 2 (call the force F_2). The field due to I_1 at a distance *r* is

Equation:

$$B_1 = \frac{\mu_0 I_1}{2\pi r}$$



(a) The magnetic field produced by a long straight conductor is perpendicular to a parallel conductor, as indicated by right-hand rule (RHR)-2. (b) A view from above of the two wires shown in (a), with one magnetic field line shown for wire 1. RHR-1 shows that the force between the parallel conductors is attractive when the currents are in the same direction. A similar analysis shows that the force is repulsive between currents in opposite directions.

This field is uniform from the wire 1 and perpendicular to it, so the force F_2 it exerts on a length *l* of wire 2 is given by $F = IlB \sin \theta$ with $\sin \theta = 1$:

Equation:

$$F_2 = I_2 l B_1.$$

The forces on the wires are equal in magnitude, so we just write *F* for the magnitude of F_2. (Note that $\vec{\mathbf{F}}_1 = -\vec{\mathbf{F}}_2$.) Since the wires are very long, it is convenient to think in terms of *F*/*l*, the force per unit length. Substituting the expression for B_1 into [link] and rearranging terms gives

Note:

Equation:

$$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}.$$

The ratio *F*/*l* is the force per unit length between two parallel currents I_1 and I_2 separated by a distance *r*. The force is attractive if the currents are in the same direction and repulsive if they are in opposite directions.

This force is responsible for the *pinch effect* in electric arcs and other plasmas. The force exists whether the currents are in wires or not. It is only apparent if the overall charge density is zero; otherwise, the Coulomb repulsion overwhelms the magnetic attraction. In an electric arc, where charges are moving parallel to one another, an attractive force squeezes currents into a smaller tube. In large circuit breakers, such as those used in neighborhood power distribution systems, the pinch effect can concentrate an arc between plates of a switch trying to break a large current, burn holes, and even ignite the equipment. Another example of the pinch effect is found in the solar plasma, where jets of ionized material, such as solar flares, are shaped by magnetic forces.

The definition of the ampere is based on the force between current-carrying wires. Note that for long, parallel wires separated by 1 meter with each carrying 1 ampere, the force per meter is

Equation:

$$\frac{F}{l} = \frac{\left(4\pi \times 10^{-7}\text{T}\cdot\text{m/A}\right)(1\text{ A})^2}{(2\pi)(1\text{ m})} = 2 \times 10^{-7}\text{ N/m.}$$

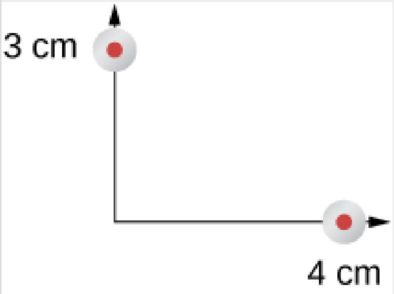
Since μ_0 is exactly $4\pi \times 10^{-7}\text{ T}\cdot\text{m/A}$ by definition, and because $1\text{ T} = 1\text{ N/(A}\cdot\text{m)}$, the force per meter is exactly $2 \times 10^{-7}\text{ N/m}$. This is the basis of the definition of the ampere.

Infinite-length wires are impractical, so in practice, a current balance is constructed with coils of wire separated by a few centimeters. Force is measured to determine current. This also provides us with a method for measuring the coulomb. We measure the charge that flows for a current of one ampere in one second. That is, $1\text{ C} = 1\text{ A}\cdot\text{s}$. For both the ampere and the coulomb, the method of measuring force between conductors is the most accurate in practice.

Example:

Calculating Forces on Wires

Two wires, both carrying current out of the page, have a current of magnitude 5.0 mA. The first wire is located at (0.0 cm, 3.0 cm) while the other wire is located at (4.0 cm, 0.0 cm) as shown in [link]. What is the magnetic force per unit length of the first wire on the second and the second wire on the first?



Two current-carrying wires at given locations with currents out of the page.

Strategy

Each wire produces a magnetic field felt by the other wire. The distance along the hypotenuse of the triangle between the wires is the radial distance used in the calculation to determine the force per unit length. Since both wires have currents flowing in the same direction, the direction of the force is toward each other.

Solution

The distance between the wires results from finding the hypotenuse of a triangle:

Equation:

$$r=\sqrt{(3.0\text{ cm})^2+(4.0\text{ cm})^2}=5.0\text{ cm}.$$

The force per unit length can then be calculated using the known currents in the wires:

Equation:

$$\frac{F}{l}=\frac{\left(4\pi\times10^{-7}\text{T}\cdot\text{m/A}\right)\left(5\times10^{-3}\text{A}\right)^2}{(2\pi)(5\times10^{-2}\text{m})}=1\times10^{-10}\,\text{N/m}.$$

The force from the first wire pulls the second wire. The angle between the radius and the *x*-axis is

Equation:

$$\theta=\tan^{-1}\left(\frac{3\text{ cm}}{4\text{ cm}}\right)=36.9^{\circ}.$$

The unit vector for this is calculated by

Equation:

$$\cos(36.9^{\circ})\hat{\mathbf{i}}-\sin(36.9^{\circ})\hat{\mathbf{j}}=0.8\hat{\mathbf{i}}-0.6\hat{\mathbf{j}}.$$

Therefore, the force per unit length from wire one on wire 2 is

Equation:

$$\frac{\vec{\mathbf{F}}}{l} = (1 \times 10^{-10}\,\text{N/m}) \times (0.8\hat{\mathbf{i}} - 0.6\hat{\mathbf{j}}) = (8 \times 10^{-11}\hat{\mathbf{i}} - 6 \times 10^{-11}\hat{\mathbf{j}})\,\text{N/m}.$$

The force per unit length from wire 2 on wire 1 is the negative of the previous answer:

Equation:

$$\frac{\vec{\mathbf{F}}}{l} = (-8 \times 10^{-11}\hat{\mathbf{i}} + 6 \times 10^{-11}\hat{\mathbf{j}})\text{N/m}.$$

Significance

These wires produced magnetic fields of equal magnitude but opposite directions at each other's locations. Whether the fields are identical or not, the forces that the wires exert on each other are always equal in magnitude and opposite in direction (Newton's third law).

Note:

Exercise:

Problem:

Check Your Understanding Two wires, both carrying current out of the page, have a current of magnitude 2.0 mA and 3.0 mA, respectively. The first wire is located at (0.0 cm, 5.0 cm) while the other wire is located at (12.0 cm, 0.0 cm). What is the magnitude of the magnetic force per unit length of the first wire on the second and the second wire on the first?

Solution:

Both have a force per unit length of $9.23 \times 10^{-12}\,\text{N/m}$

Summary

- The force between two parallel currents I_1 and I_2, separated by a distance r, has a magnitude per unit length given by $\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}$.
- The force is attractive if the currents are in the same direction, repulsive if they are in opposite directions.

Conceptual Questions

Exercise:

Problem:

Compare and contrast the electric field of an infinite line of charge and the magnetic field of an infinite line of current.

Exercise:

Problem:

Is $\vec{\mathbf{B}}$ constant in magnitude for points that lie on a magnetic field line?

Solution:

A magnetic field line gives the direction of the magnetic field at any point in space. The density of magnetic field lines indicates the strength of the magnetic field.

Problems

Exercise:

Problem:

Two long, straight wires are parallel and 25 cm apart. (a) If each wire carries a current of 50 A in the same direction, what is the magnetic force per meter exerted on each wire? (b) Does the force pull the wires together or push them apart? (c) What happens if the currents flow in opposite directions?

Exercise:

Problem:

Two long, straight wires are parallel and 10 cm apart. One carries a current of 2.0 A, the other a current of 5.0 A. (a) If the two currents flow in opposite directions, what is the magnitude and direction of the force per unit length of one wire on the other? (b) What is the magnitude and direction of the force per unit length if the currents flow in the same direction?

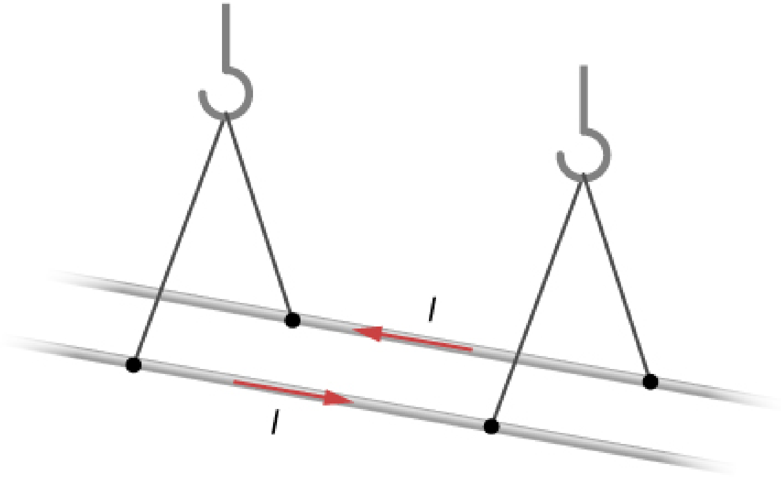
Solution:

a. $F/l = 8\ \times\ 10^{-6}\text{N/m}$ away from the other wire; b. $F/l = 8\ \times\ 10^{-6}\text{N/m}$ toward the other wire

Exercise:

Problem:

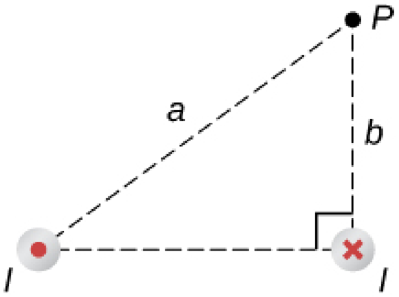
Two long, parallel wires are hung by cords of length 5.0 cm, as shown in the accompanying figure. Each wire has a mass per unit length of 30 g/m, and they carry the same current in opposite directions. What is the current if the cords hang at 6.0° with respect to the vertical?



Exercise:

Problem:

A circuit with current *I* has two long parallel wire sections that carry current in opposite directions. Find magnetic field at a point *P* near these wires that is a distance *a* from one wire and *b* from the other wire as shown in the figure.



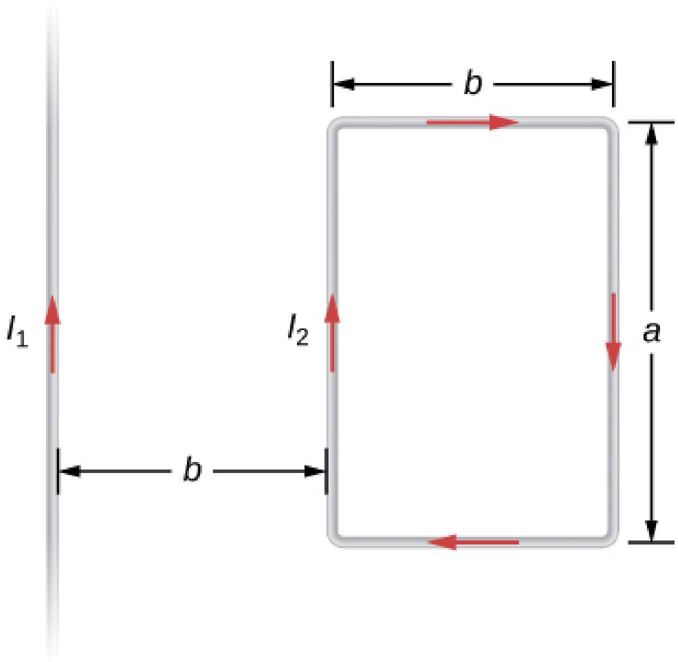
Solution:

$$B = \frac{\mu_o I}{2\pi a^2 b}\left(\left(a_2 + b_2\right)\hat{\mathbf{i}} + b\sqrt{(a^2 - b^2)}\hat{\mathbf{j}}\right)$$

Exercise:

Problem:

The infinite, straight wire shown in the accompanying figure carries a current I_1. The rectangular loop, whose long sides are parallel to the wire, carries a current I_2. What are the magnitude and direction of the force on the rectangular loop due to the magnetic field of the wire?

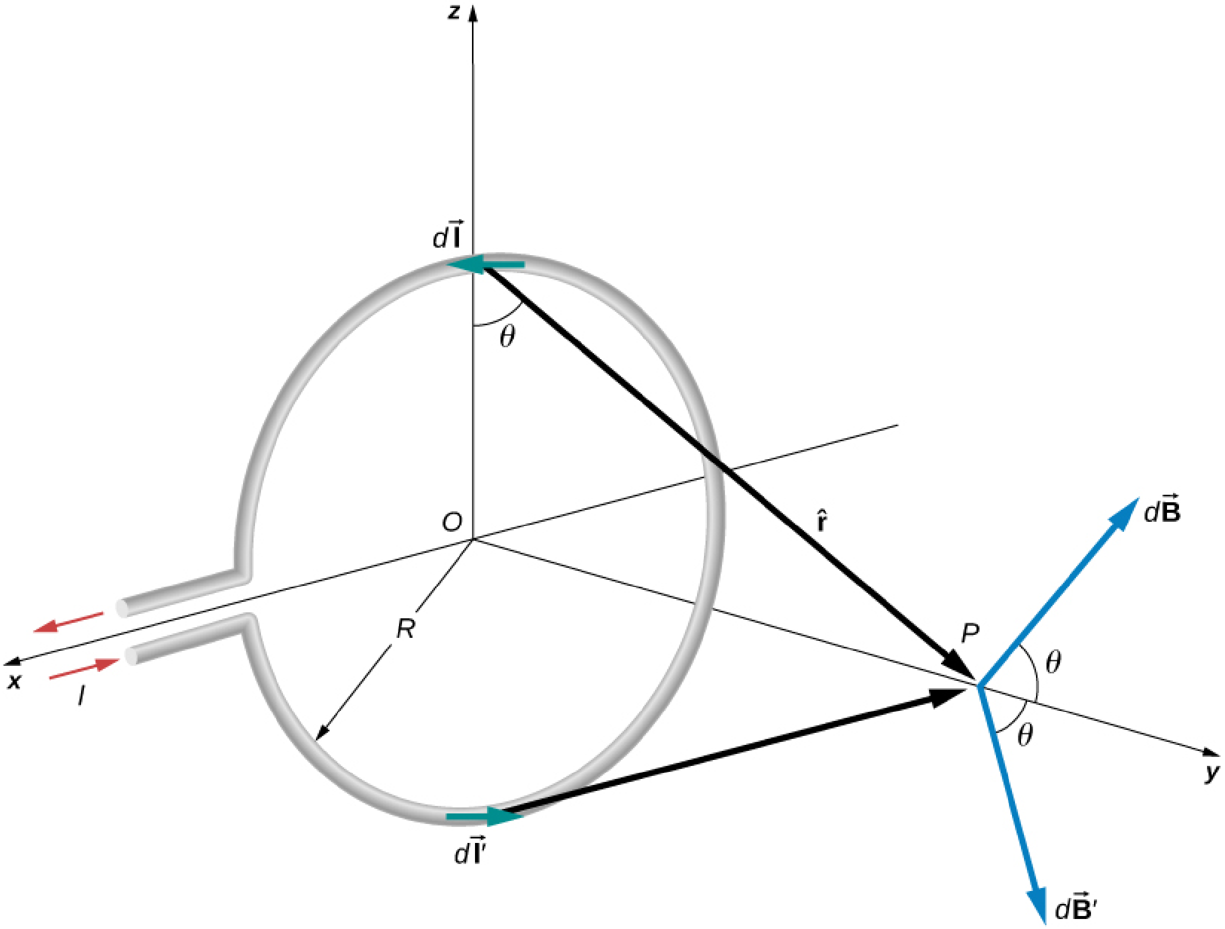


Magnetic Field of a Current Loop

By the end of this section, you will be able to:

- Explain how the Biot-Savart law is used to determine the magnetic field due to a current in a loop of wire at a point along a line perpendicular to the plane of the loop.
- Determine the magnetic field of an arc of current.

The circular loop of [link] has a radius *R*, carries a current *I*, and lies in the *xz*-plane. What is the magnetic field due to the current at an arbitrary point *P* along the axis of the loop?



z
$d\vec{\mathbf{l}}$
θ
O
$\hat{\mathbf{r}}$
$d\vec{\mathbf{B}}$
P
θ
θ
R
x
I
y
$d\vec{\mathbf{l}}'$
$d\vec{\mathbf{B}}'$

Determining the magnetic field at point *P* along the axis of a current-carrying loop of wire.

We can use the Biot-Savart law to find the magnetic field due to a current. We first consider arbitrary segments on opposite sides of the loop to qualitatively show by the vector results that the net magnetic field direction is along the central axis from the loop. From there, we can use the Biot-Savart law to derive the expression for magnetic field.

Let *P* be a distance *y* from the center of the loop. From the right-hand rule, the magnetic field $d\vec{\mathbf{B}}$ at *P*, produced by the current element $I\,d\vec{\mathbf{l}}$, is directed at an angle *θ* above the *y*-axis as shown. Since $d\vec{\mathbf{l}}$ is parallel along the *x*-axis and $\hat{\mathbf{r}}$ is in the *yz*-plane, the two vectors are perpendicular, so we have

Equation:

$$dB = \frac{\mu_0}{4\pi}\,\frac{I\,dl\,\sin\pi/2}{r^2} = \frac{\mu_0}{4\pi}\,\frac{I\,dl}{y^2+R^2}$$

where we have used $r^2 = y^2 + R^2$.

Now consider the magnetic field $d\vec{\mathbf{B}}'$ due to the current element $I\,d\vec{\mathbf{l}}'$, which is directly opposite $I\,d\vec{\mathbf{l}}$ on the loop. The magnitude of $d\vec{\mathbf{B}}'$ is also given by [link], but it is directed at an angle *θ* below the *y*-axis. The components of $d\vec{\mathbf{B}}$ and $d\vec{\mathbf{B}}'$ perpendicular to the *y*-axis therefore cancel, and in calculating the net magnetic field, only the components along the *y*-axis need to be considered. The components perpendicular to the axis of the loop sum to zero in pairs. Hence at point *P*:

Equation:

$$\vec{\mathbf{B}} = \hat{\mathbf{j}}\int_{\text{loop}} dB\,\cos\theta = \hat{\mathbf{j}}\frac{\mu_0 I}{4\pi}\int_{\text{loop}} \frac{\cos\theta\,dl}{y^2+R^2}.$$

For all elements $d\vec{\mathbf{l}}$ on the wire, *y*, *R*, and cos *θ* are constant and are related by

Equation:

$$\cos\theta = \frac{R}{\sqrt{y^2+R^2}}.$$

Now from [link], the magnetic field at P is

Equation:

$$\vec{\mathbf{B}} = \hat{\mathbf{j}}\frac{\mu_0 IR}{4\pi(y^2+R^2)^{3/2}}\int_{\text{loop}} dl = \frac{\mu_0 IR^2}{2(y^2+R^2)^{3/2}}\hat{\mathbf{j}}$$

where we have used $\int_{\text{loop}} dl = 2\pi R$. As discussed in the previous chapter, the closed current loop is a magnetic dipole of moment $\vec{\mu} = IA\widehat{\mathbf{n}}$. For this example, $A = \pi R^2$ and $\widehat{\mathbf{n}} = \hat{\mathbf{j}}$, so the magnetic field at P can also be written as

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0 \mu \hat{\mathbf{j}}}{2\pi(y^2+R^2)^{3/2}}.$$

By setting $y = 0$ in [link], we obtain the magnetic field at the center of the loop:

Note:

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0 I}{2R}\hat{\mathbf{j}}.$$

This equation becomes $B = \mu_0 nI/(2R)$ for a flat coil of *n* loops per length. It can also be expressed as

Equation:

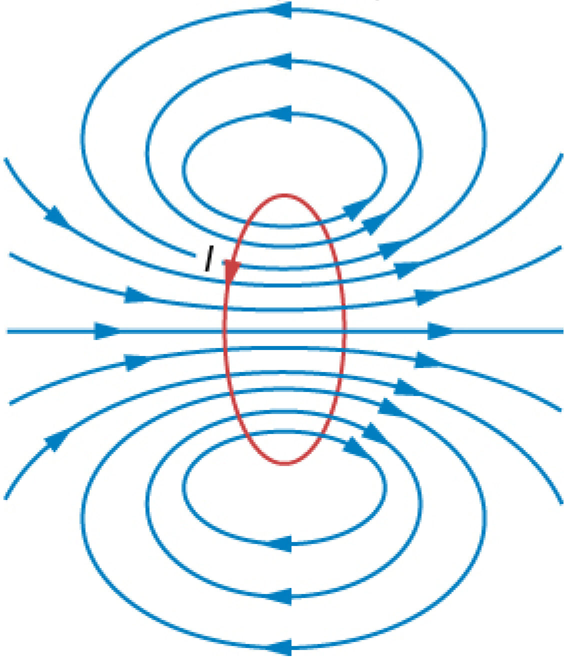
$$\vec{\mathbf{B}} = \frac{\mu_0 \vec{\mu}}{2\pi R^3}.$$

If we consider $y \gg R$ in [link], the expression reduces to an expression known as the magnetic field from a dipole:

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0 \vec{\mu}}{2\pi y^3}.$$

The calculation of the magnetic field due to the circular current loop at points off-axis requires rather complex mathematics, so we'll just look at the results. The magnetic field lines are shaped as shown in [link]. Notice that one field line follows the axis of the loop. This is the field line we just found. Also, very close to the wire, the field lines are almost circular, like the lines of a long straight wire.

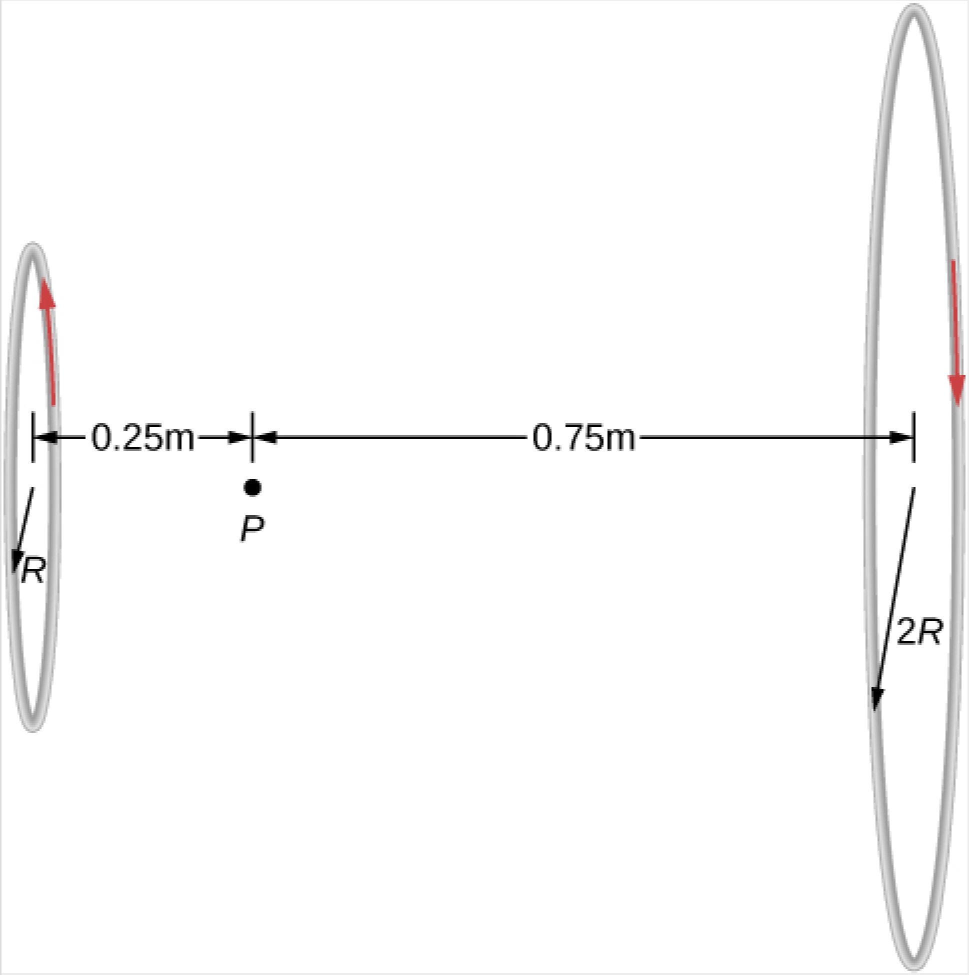


Sketch of the magnetic field lines of a circular current loop.

Example:

Magnetic Field between Two Loops

Two loops of wire carry the same current of 10 mA, but flow in opposite directions as seen in [link]. One loop is measured to have a radius of $R = 50\text{ cm}$ while the other loop has a radius of $2R = 100\text{ cm}$. The distance from the first loop to the point where the magnetic field is measured is 0.25 m, and the distance from that point to the second loop is 0.75 m. What is the magnitude of the net magnetic field at point *P*?



Two loops of different radii have the same current but flowing in opposite directions. The magnetic field at point *P* is measured to be zero.

Strategy

The magnetic field at point *P* has been determined in [link]. Since the currents are flowing in opposite directions, the net magnetic field is the difference between the two fields generated by the coils. Using the given quantities in the problem, the net magnetic field is then calculated.

Solution

Solving for the net magnetic field using [link] and the given quantities in the problem yields

Equation:

$$\begin{aligned}
B &= \frac{\mu_0 I {R_1}^2}{2({y_1}^2+{R_1}^2)^{3/2}} - \frac{\mu_0 I {R_2}^2}{2({y_2}^2+{R_2}^2)^{3/2}} \\
B &= \frac{(4\pi \times 10^{-7}\,\text{T}\cdot\text{m/A})(0.010\,\text{A})(0.5\,\text{m})^2}{2((0.25\,\text{m})^2+(0.5\,\text{m})^2)^{3/2}} - \frac{(4\pi \times 10^{-7}\,\text{T}\cdot\text{m/A})(0.010\,\text{A})(1.0\,\text{m})^2}{2((0.75\,\text{m})^2+(1.0\,\text{m})^2)^{3/2}} \\
B &= 5.77 \times 10^{-9}\,\text{T to the right.}
\end{aligned}$$

Significance

Helmholtz coils typically have loops with equal radii with current flowing in the same direction to have a strong uniform field at the midpoint between the loops. A similar application of the magnetic field distribution created by Helmholtz coils is found in a magnetic bottle that can temporarily trap charged particles. See Magnetic Forces and Fields for a discussion on this.

Note:

Exercise:

Problem:

Check Your Understanding Using [link], at what distance would you have to move the first coil to have zero measurable magnetic field at point P?

Solution:

0.608 meters

Summary

- The magnetic field strength at the center of a circular loop is given by $B = \frac{\mu_0 I}{2R} \text{ (at center of loop)}$, where R is the radius of the loop. RHR-2 gives the direction of the field about the loop.

Conceptual Questions

Exercise:

Problem: Is the magnetic field of a current loop uniform?

Exercise:

Problem:

What happens to the length of a suspended spring when a current passes through it?

Solution:

The spring reduces in length since each coil with have a north pole-produced magnetic field next to a south pole of the next coil.

Exercise:

Problem:

Two concentric circular wires with different diameters carry currents in the same direction. Describe the force on the inner wire.

Problems

Exercise:

Problem:

When the current through a circular loop is 6.0 A, the magnetic field at its center is $2.0 \times 10^{-4} \mathrm{T}$. What is the radius of the loop?

Solution:

0.019 m

Exercise:

Problem:

How many turns must be wound on a flat, circular coil of radius 20 cm in order to produce a magnetic field of magnitude $4.0 \times 10^{-5} \mathrm{T}$ at the center of the coil when the current through it is 0.85 A?

Exercise:

Problem:

A flat, circular loop has 20 turns. The radius of the loop is 10.0 cm and the current through the wire is 0.50 A. Determine the magnitude of the magnetic field at the center of the loop.

Solution:

$\mathrm{N} \times 6.28 \times 10^{-5} \mathrm{T}$

Exercise:

Problem:

A circular loop of radius *R* carries a current *I*. At what distance along the axis of the loop is the magnetic field one-half its value at the center of the loop?

Exercise:

Problem:

Two flat, circular coils, each with a radius *R* and wound with *N* turns, are mounted along the same axis so that they are parallel a distance *d* apart. What is the magnetic field at the midpoint of the common axis if a current *I* flows in the same direction through each coil?

Solution:

$$B=\frac{\mu_o I R^2}{\left(\left(\frac{d}{2}\right)^2+R^2\right)^{3/2}}$$

Exercise:

Problem:

For the coils in the preceding problem, what is the magnetic field at the center of either coil?

Ampère’s Law

By the end of this section, you will be able to:

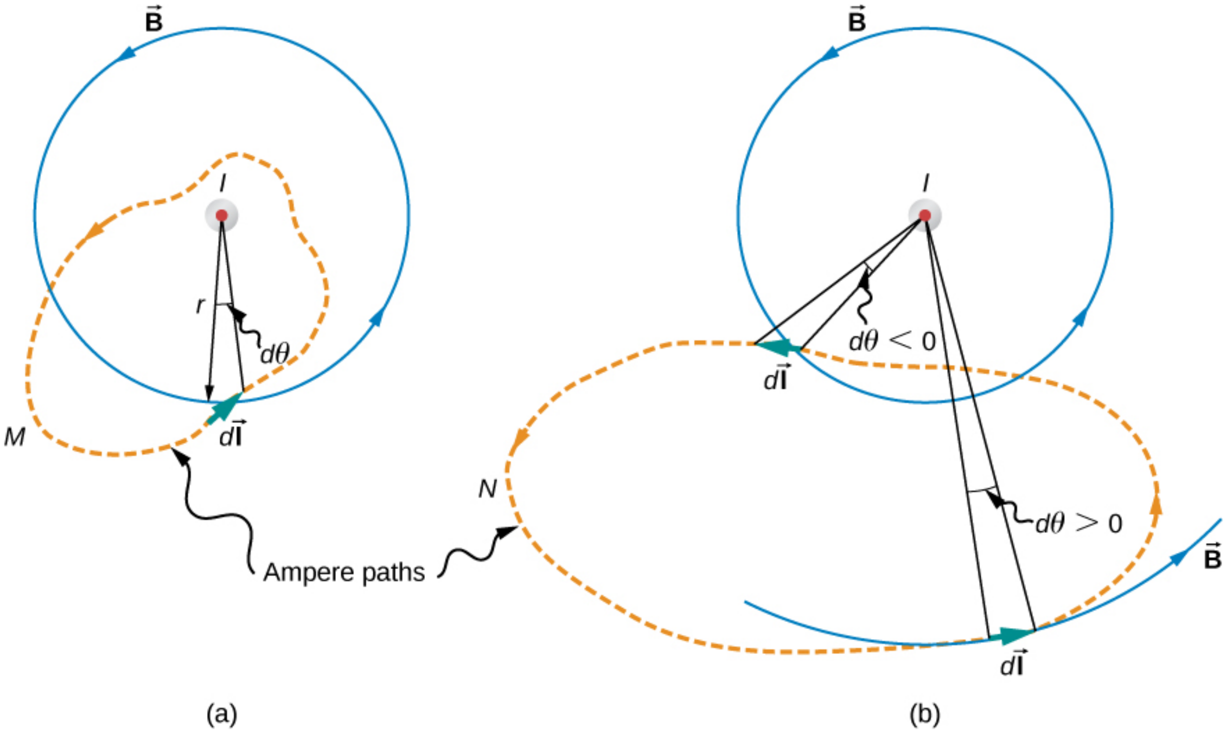
- Explain how Ampère’s law relates the magnetic field produced by a current to the value of the current
- Calculate the magnetic field from a long straight wire, either thin or thick, by Ampère’s law

A fundamental property of a static magnetic field is that, unlike an electrostatic field, it is not conservative. A conservative field is one that does the same amount of work on a particle moving between two different points regardless of the path chosen. Magnetic fields do not have such a property. Instead, there is a relationship between the magnetic field and its source, electric current. It is expressed in terms of the line integral of $\vec{\mathbf{B}}$ and is known as **Ampère’s law**. This law can also be derived directly from the Biot-Savart law. We now consider that derivation for the special case of an infinite, straight wire.

[link] shows an arbitrary plane perpendicular to an infinite, straight wire whose current I is directed out of the page. The magnetic field lines are circles directed counterclockwise and centered on the wire. To begin, let’s consider $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ over the closed paths M and N. Notice that one path (M) encloses the wire, whereas the other (N) does not. Since the field lines are circular, $\vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ is the product of B and the projection of dl onto the circle passing through $d\vec{\mathbf{l}}$. If the radius of this particular circle is r, the projection is $rd\theta$, and

Equation:

$$\vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = Br\,d\theta.$$


$\vec{\mathbf{B}}$
I
r
$d\theta$
M
$d\vec{\mathbf{l}}$
Ampere paths
(a)
$\vec{\mathbf{B}}$
I
$d\theta < 0$
$d\vec{\mathbf{l}}$
N
$d\theta > 0$
$\vec{\mathbf{B}}$
$d\vec{\mathbf{l}}$
(b)

The current I of a long, straight wire is directed out of the page. The integral $\oint d\theta$ equals 2π and 0, respectively, for paths M and N.

With $\overrightarrow{\mathbf{B}}$ given by [link],

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \oint \left(\frac{\mu_0 I}{2\pi r}\right) r\, d\theta = \frac{\mu_0 I}{2\pi} \oint d\theta.$$

For path M, which circulates around the wire, $\oint_M d\theta = 2\pi$ and

Equation:

$$\oint_M \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0 I.$$

Path N, on the other hand, circulates through both positive (counterclockwise) and negative (clockwise) $d\theta$ (see [link]), and since it is closed, $\oint_N d\theta = 0$. Thus for path N,

Equation:

$$\oint_N \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = 0.$$

The extension of this result to the general case is Ampère’s law.

Note:

Ampère’s law

Over an arbitrary closed path,

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0 I$$

where I is the total current passing through any open surface S whose perimeter is the path of integration. Only currents inside the path of integration need be considered.

To determine whether a specific current I is positive or negative, curl the fingers of your right hand in the direction of the path of integration, as shown in [link]. If I passes through S in the same direction as your extended thumb, I is positive; if I passes through S in the direction opposite to your extended thumb, it is negative.

Note:

Problem-Solving Strategy: Ampère’s Law

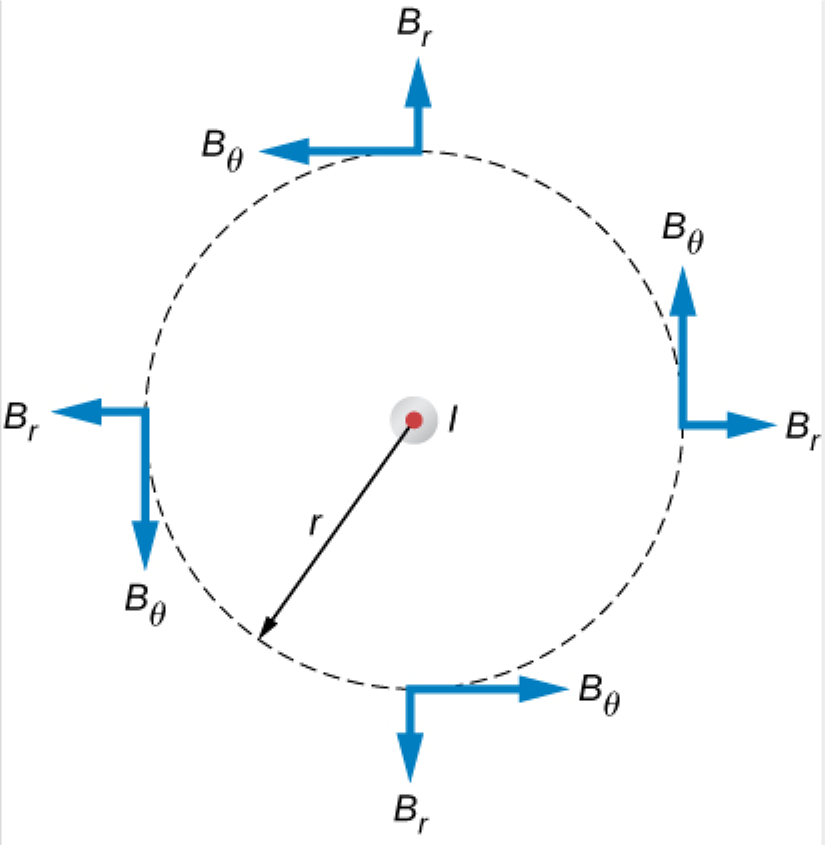
To calculate the magnetic field created from current in wire(s), use the following steps:

1. Identify the symmetry of the current in the wire(s). If there is no symmetry, use the Biot-Savart law to determine the magnetic field.
2. Determine the direction of the magnetic field created by the wire(s) by right-hand rule 2.
3. Chose a path loop where the magnetic field is either constant or zero.
4. Calculate the current inside the loop.
5. Calculate the line integral $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ around the closed loop.
6. Equate $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ with $\mu_0 I_{\text{enc}}$ and solve for $\vec{\mathbf{B}}$.

Example:

Using Ampère’s Law to Calculate the Magnetic Field Due to a Wire

Use Ampère’s law to calculate the magnetic field due to a steady current I in an infinitely long, thin, straight wire as shown in [link].



The possible components of the magnetic field *B* due to a current *I*, which is directed out of the page. The radial component is zero because the angle between the magnetic field and the path is at a right angle.

Strategy
Consider an arbitrary plane perpendicular to the wire, with the current directed out of the page. The possible magnetic field components in this plane, B_r and B_θ, are shown at arbitrary points on a circle of radius r centered on the wire. Since the field is cylindrically symmetric, neither B_r nor B_θ varies with the position on this circle. Also from symmetry, the radial lines, if they exist, must be directed either all inward or all outward from the wire. This means, however, that there must be a net magnetic flux across an arbitrary cylinder concentric with the wire. The radial component

of the magnetic field must be zero because $\vec{\mathbf{B}}_r \cdot d\vec{\mathbf{l}} = 0$. Therefore, we can apply Ampère’s law to the circular path as shown.

Solution

Over this path $\vec{\mathbf{B}}$ is constant and parallel to $d\vec{\mathbf{l}}$, so

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = B_\theta \oint dl = B_\theta(2\pi r).$$

Thus Ampère’s law reduces to

Equation:

$$B_\theta(2\pi r) = \mu_0 I.$$

Finally, since B_θ is the only component of $\vec{\mathbf{B}}$, we can drop the subscript and write

Equation:

$$B = \frac{\mu_0 I}{2\pi r}.$$

This agrees with the Biot-Savart calculation above.

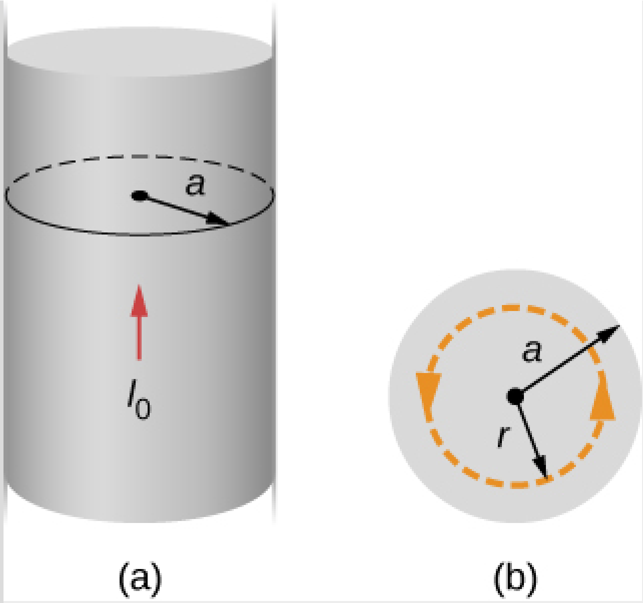
Significance

Ampère’s law works well if you have a path to integrate over which $\vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ has results that are easy to simplify. For the infinite wire, this works easily with a path that is circular around the wire so that the magnetic field factors out of the integration. If the path dependence looks complicated, you can always go back to the Biot-Savart law and use that to find the magnetic field.

Example:

Calculating the Magnetic Field of a Thick Wire with Ampère’s Law

The radius of the long, straight wire of [link] is *a*, and the wire carries a current I_0 that is distributed uniformly over its cross-section. Find the magnetic field both inside and outside the wire.



(a) A model of a current-carrying wire of radius a and current I_0. (b) A cross-section of the same wire showing the radius a and the Ampère’s loop of radius r.

Strategy

This problem has the same geometry as [link], but the enclosed current changes as we move the integration path from outside the wire to inside the wire, where it doesn’t capture the entire current enclosed (see [link]).

Solution

For any circular path of radius r that is centered on the wire,

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \oint B dl = B \oint dl = B(2\pi r).$$

From Ampère’s law, this equals the total current passing through any surface bounded by the path of integration.

Consider first a circular path that is inside the wire $(r \leq a)$ such as that shown in part (a) of [link]. We need the current I passing through the area enclosed by the path. It’s equal to the current density J times the area enclosed. Since the current is uniform, the current density inside the path equals the current density in the whole wire, which is $I_0/\pi a^2$. Therefore the current *I* passing through the area enclosed by the path is

Equation:

$$I = \frac{\pi r^2}{\pi a^2} I_0 = \frac{r^2}{a^2} I_0.$$

We can consider this ratio because the current density *J* is constant over the area of the wire. Therefore, the current density of a part of the wire is equal to the current density in the whole area. Using Ampère's law, we obtain

Equation:

$$B(2\pi r) = \mu_0 \left(\frac{r^2}{a^2}\right) I_0,$$

and the magnetic field inside the wire is

Equation:

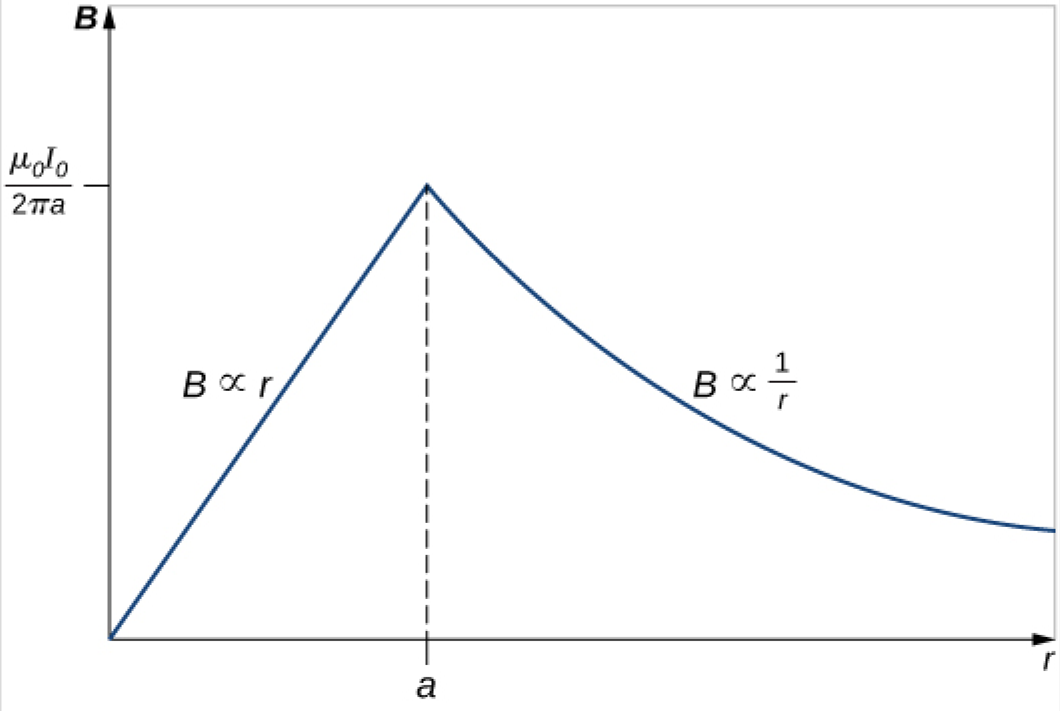
$$B = \frac{\mu_0 I_0}{2\pi} \frac{r}{a^2} \; (r \leq a).$$

Outside the wire, the situation is identical to that of the infinite thin wire of the previous example; that is,

Equation:

$$B = \frac{\mu_0 I_0}{2\pi r} \; (r \geq a).$$

The variation of *B* with *r* is shown in [link].



Variation of the magnetic field produced by a current I_0 in a long, straight wire of radius *a*.

Significance

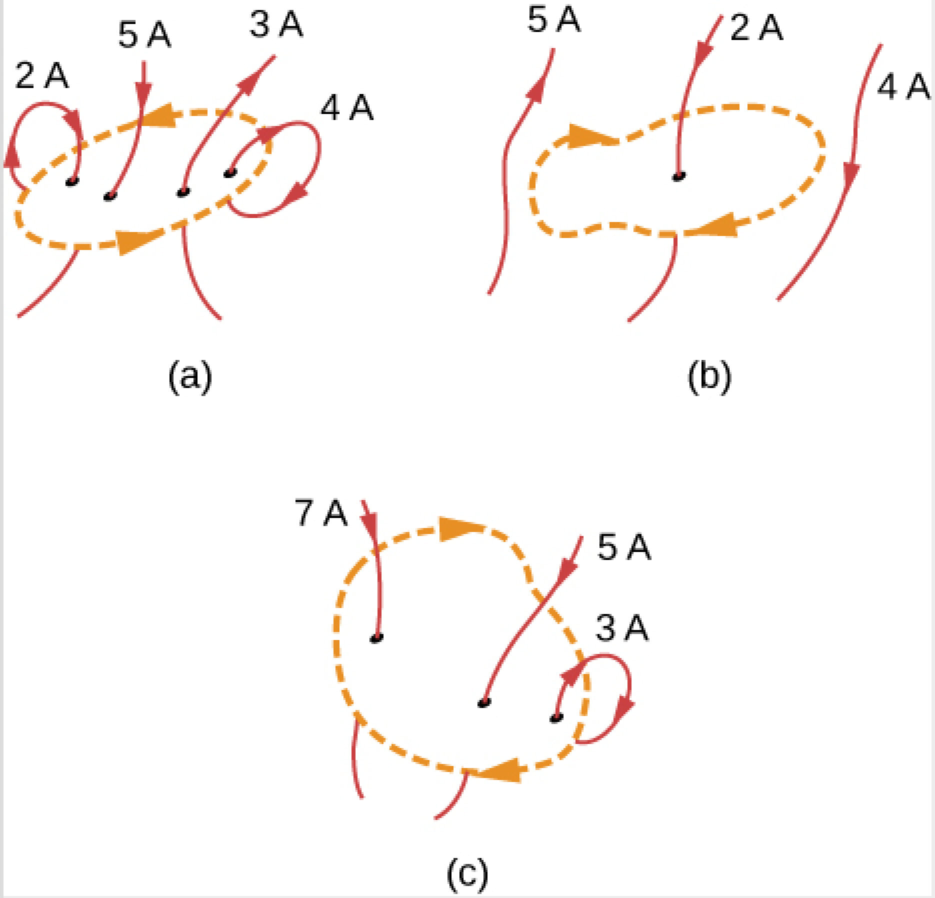
The results show that as the radial distance increases inside the thick wire, the magnetic field increases from zero to a familiar value of the magnetic field of a thin wire. Outside the wire, the field drops off regardless of whether it was a thick or thin wire.

This result is similar to how Gauss's law for electrical charges behaves inside a uniform charge distribution, except that Gauss's law for electrical charges has a uniform volume distribution of charge, whereas Ampère's law here has a uniform area of current distribution. Also, the drop-off outside the thick wire is similar to how an electric field drops off outside of a linear charge distribution, since the two cases have the same geometry and neither case depends on the configuration of charges or currents once the loop is outside the distribution.

Example:

Using Ampère’s Law with Arbitrary Paths

Use Ampère’s law to evaluate $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ for the current configurations and paths in [link].



Current configurations and paths for [link].

Strategy

Ampère’s law states that $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0 I$ where I is the total current passing through the enclosed loop. The quickest way to evaluate the integral is to calculate $\mu_0 I$ by finding the net current through the loop. Positive currents flow with your right-hand thumb if your fingers wrap around in the direction of the loop. This will tell us the sign of the answer.

Solution

(a) The current going downward through the loop equals the current going out of the loop, so the net current is zero. Thus, $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = 0.$

(b) The only current to consider in this problem is 2A because it is the only current inside the loop. The right-hand rule shows us the current going downward through the loop is in the positive direction. Therefore, the answer is $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0(2\text{ A}) = 2.51 \times 10^{-6}\text{T} \cdot \text{m}.$

(c) The right-hand rule shows us the current going downward through the loop is in the positive direction. There are $7\text{A} + 5\text{A} = 12\text{A}$ of current going downward and –3 A going upward. Therefore, the total current is 9 A and $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0(9\text{ A}) = 5.65 \times 10^{-6}\text{T} \cdot \text{m}.$

Significance

If the currents all wrapped around so that the same current went into the loop and out of the loop, the net current would be zero and no magnetic field would be present. This is why wires are very close to each other in an electrical cord. The currents flowing toward a device and away from a device in a wire equal zero total current flow through an Ampère loop around these wires. Therefore, no stray magnetic fields can be present from cords carrying current.

Note:

Exercise:

Problem:

Check Your Understanding Consider using Ampère’s law to calculate the magnetic fields of a finite straight wire and of a circular loop of wire. Why is it not useful for these calculations?

Solution:

In these cases the integrals around the Ampèrian loop are very difficult because there is no symmetry, so this method would not be useful.

Summary

- The magnetic field created by current following any path is the sum (or integral) of the fields due to segments along the path (magnitude and direction as for a straight wire), resulting in a general relationship between current and field known as Ampère’s law.
- Ampère’s law can be used to determine the magnetic field from a thin wire or thick wire by a geometrically convenient path of integration. The results are consistent with the Biot-Savart law.

Conceptual Questions

Exercise:

Problem:

Is Ampère’s law valid for all closed paths? Why isn’t it normally useful for calculating a magnetic field?

Solution:

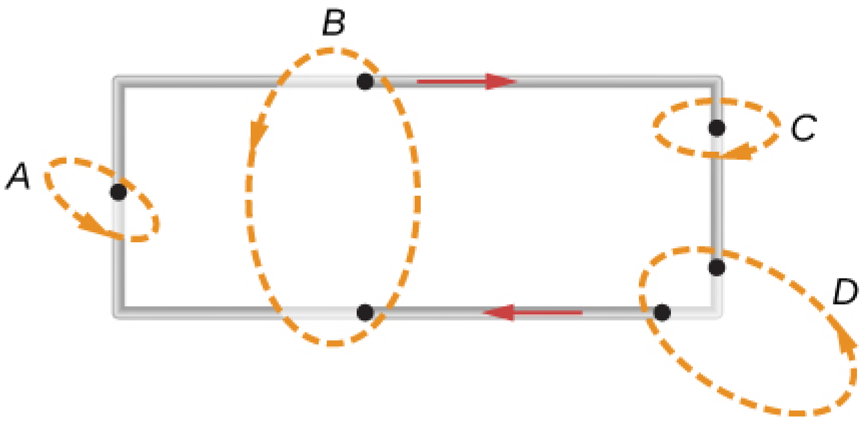
Ampère’s law is valid for all closed paths, but it is not useful for calculating fields when the magnetic field produced lacks symmetry that can be exploited by a suitable choice of path.

Problems

Exercise:

Problem:

A current I flows around the rectangular loop shown in the accompanying figure. Evaluate $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ for the paths A, B, C, and D.



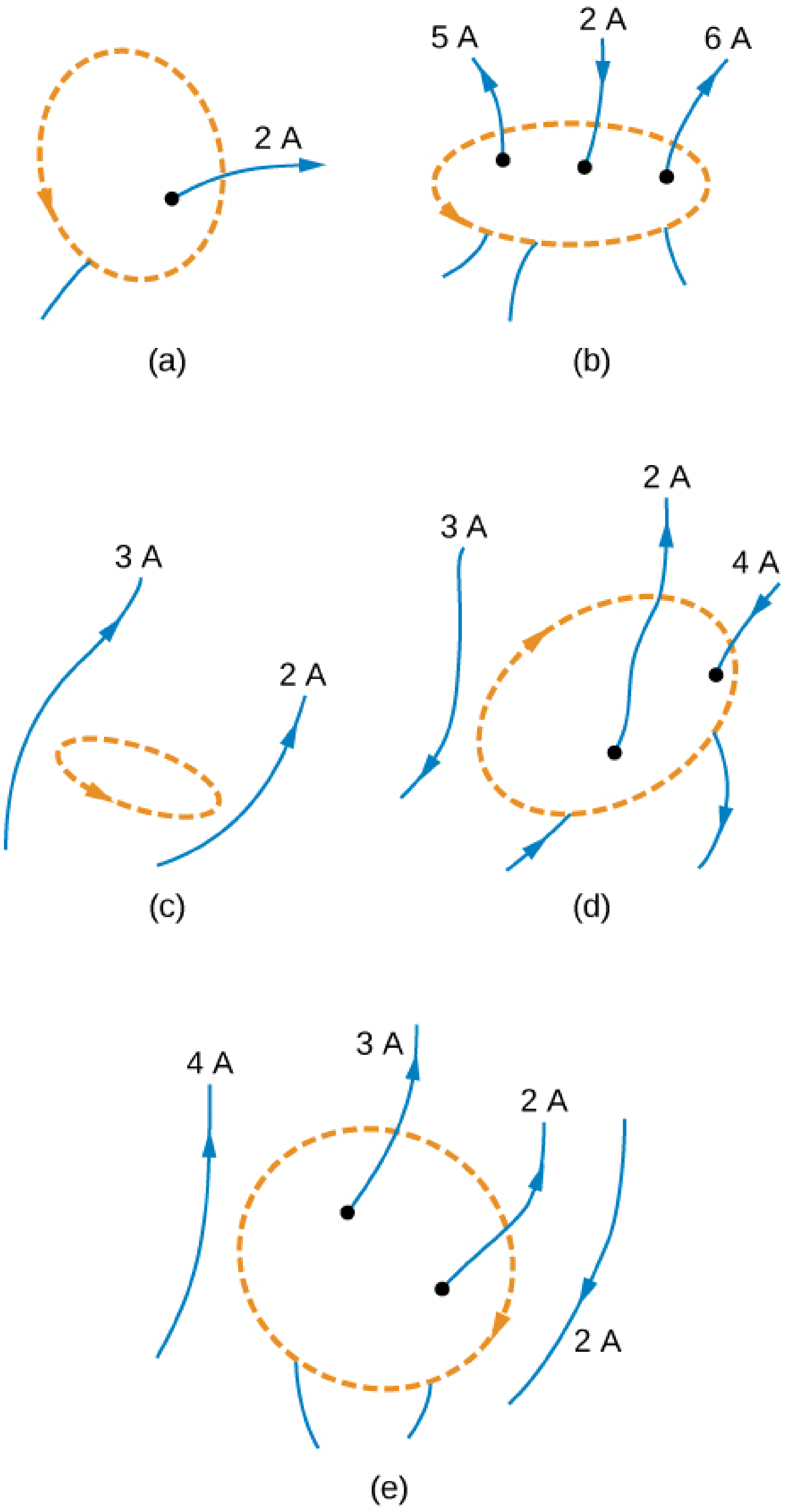
Solution:

a. $\mu_0 I$; b. 0; c. $\mu_0 I$; d. 0

Exercise:

Problem:

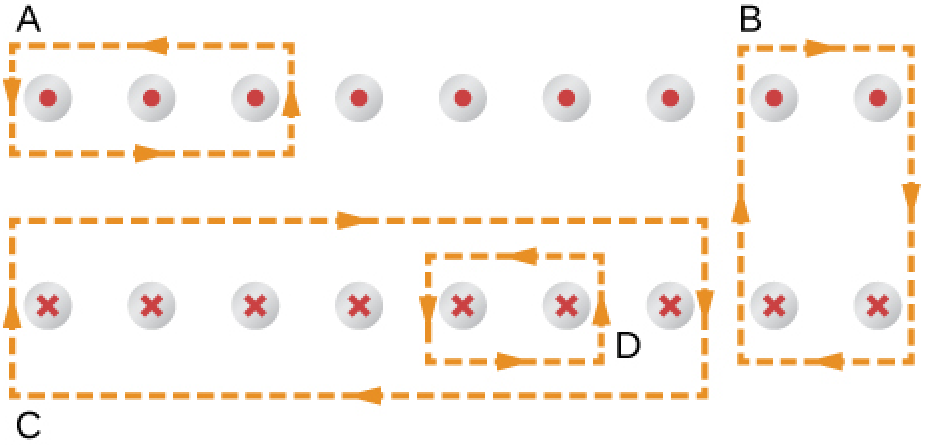
Evaluate $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ for each of the cases shown in the accompanying figure.


2 A
(a)
5 A
2 A
6 A
(b)
3 A
2 A
(c)
3 A
2 A
4 A
(d)
4 A
3 A
2 A
2 A
(e)

Exercise:

Problem:

The coil whose lengthwise cross section is shown in the accompanying figure carries a current *I* and has *N* evenly spaced turns distributed along the length l. Evaluate $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ for the paths indicated.



Solution:

a. $3\mu_0 I$; b. 0; c. $7\mu_0 I$; d. $-2\mu_0 I$

Exercise:

Problem:

A superconducting wire of diameter 0.25 cm carries a current of 1000 A. What is the magnetic field just outside the wire?

Exercise:

Problem:

A long, straight wire of radius *R* carries a current *I* that is distributed uniformly over the cross-section of the wire. At what distance from the axis of the wire is the magnitude of the magnetic field a maximum?

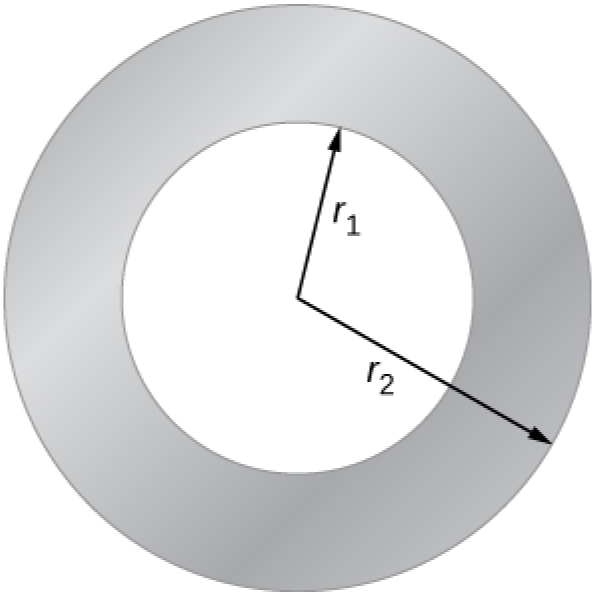
Solution:

at the radius *R*

Exercise:

Problem:

The accompanying figure shows a cross-section of a long, hollow, cylindrical conductor of inner radius $r_1 = 3.0 \text{ cm}$ and outer radius $r_2 = 5.0 \text{ cm}$. A 50-A current distributed uniformly over the cross-section flows into the page. Calculate the magnetic field at $r = 2.0 \text{ cm}, \ r = 4.0 \text{ cm}, \text{ and } r = 6.0 \text{ cm}.$

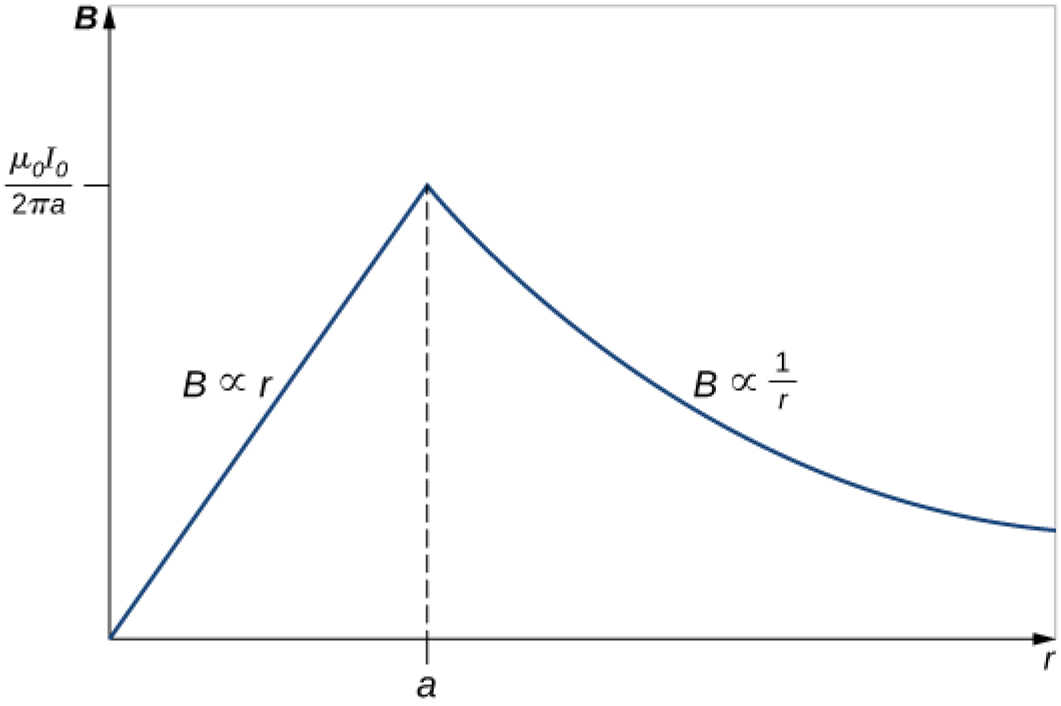


Exercise:

Problem:

A long, solid, cylindrical conductor of radius 3.0 cm carries a current of 50 A distributed uniformly over its cross-section. Plot the magnetic field as a function of the radial distance *r* from the center of the conductor.

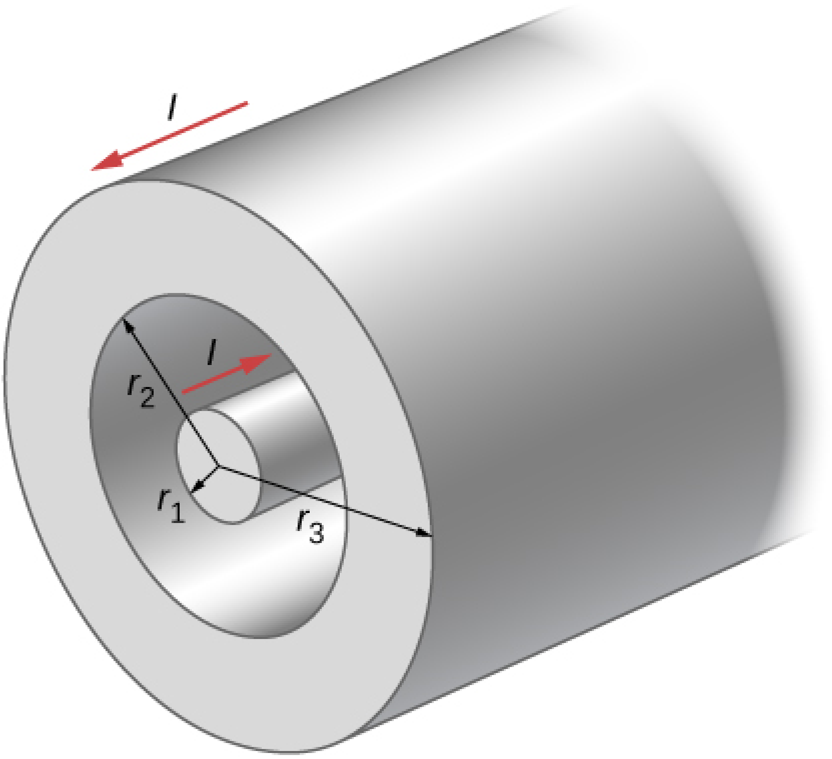
Solution:



Exercise:

Problem:

A portion of a long, cylindrical coaxial cable is shown in the accompanying figure. A current I flows down the center conductor, and this current is returned in the outer conductor. Determine the magnetic field in the regions (a) $r \le r_1$, (b) $r_2 \ge r \ge r_1$, (c) $r_3 \ge r \ge r_2$, and (d) $r \ge r_3$. Assume that the current is distributed uniformly over the cross sections of the two parts of the cable.



Glossary

Ampère’s law
    physical law that states that the line integral of the magnetic field around an electric current is proportional to the current

Solenoids and Toroids
By the end of this section, you will be able to:

- Establish a relationship for how the magnetic field of a solenoid varies with distance and current by using both the Biot-Savart law and Ampère’s law
- Establish a relationship for how the magnetic field of a toroid varies with distance and current by using Ampère’s law

Two of the most common and useful electromagnetic devices are called solenoids and toroids. In one form or another, they are part of numerous instruments, both large and small. In this section, we examine the magnetic field typical of these devices.

Solenoids

A long wire wound in the form of a helical coil is known as a **solenoid**. Solenoids are commonly used in experimental research requiring magnetic fields. A solenoid is generally easy to wind, and near its center, its magnetic field is quite uniform and directly proportional to the current in the wire.

[link] shows a solenoid consisting of N turns of wire tightly wound over a length L. A current I is flowing along the wire of the solenoid. The number of turns per unit length is N/L; therefore, the number of turns in an infinitesimal length dy are $(N/L)dy$ turns. This produces a current
Equation:

$$dI = \frac{NI}{L}dy.$$

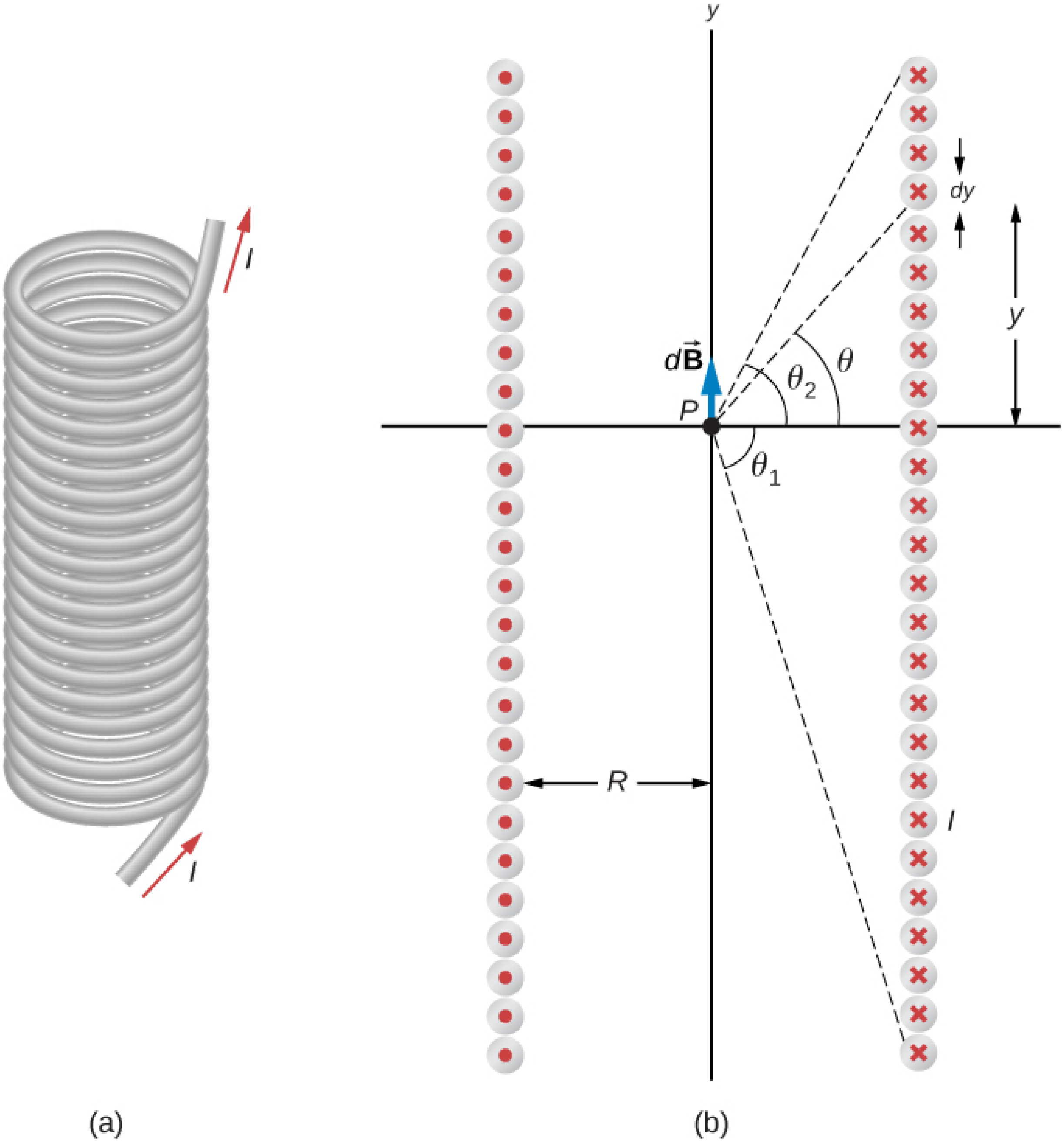
We first calculate the magnetic field at the point P of [link]. This point is on the central axis of the solenoid. We are basically cutting the solenoid into thin slices that are dy thick and treating each as a current loop. Thus, dI is the current through each slice. The magnetic field $d\vec{\mathbf{B}}$ due to the current dI in dy can be found with the help of [link] and [link]:
Equation:

$$d\vec{\mathbf{B}} = \frac{\mu_0 R^2 dI}{2(y^2 + R^2)^{3/2}}\hat{\mathbf{j}} = \left(\frac{\mu_0 I R^2 N}{2L}\hat{\mathbf{j}}\right)\frac{dy}{(y^2 + R^2)^{3/2}}$$

where we used [link] to replace dI. The resultant field at P is found by integrating $d\vec{\mathbf{B}}$ along the entire length of the solenoid. It's easiest to evaluate this integral by changing the independent variable from y to θ. From inspection of [link], we have:

Equation:

$$\sin\theta = \frac{y}{\sqrt{y^2 + R^2}}.$$



I
I
y
dy
y
θ
θ_2
$d\vec{\mathbf{B}}$
P
θ_1
R
I
(a)
(b)

(a) A solenoid is a long wire wound in the shape of a helix. (b) The magnetic field at the point *P* on the axis of the solenoid is the net field due to all of the current loops.

Taking the differential of both sides of this equation, we obtain

Equation:

$$\begin{aligned}\cos\theta\, d\theta &= \left[-\frac{y^2}{(y^2+R^2)^{3/2}} + \frac{1}{\sqrt{y^2+R^2}}\right]dy \\ &= \frac{R^2 dy}{(y^2+R^2)^{3/2}}.\end{aligned}$$

When this is substituted into the equation for $d\vec{\mathbf{B}}$, we have

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0 IN}{2L}\hat{\mathbf{j}}\int_{\theta_1}^{\theta_2}\cos\theta\, d\theta = \frac{\mu_0 IN}{2L}(\sin\theta_2 - \sin\theta_1)\hat{\mathbf{j}},$$

which is the magnetic field along the central axis of a finite solenoid.

Of special interest is the infinitely long solenoid, for which $L \to \infty$. From a practical point of view, the infinite solenoid is one whose length is much larger than its radius $(L \gg R)$. In this case, $\theta_1 = \frac{-\pi}{2}$ and $\theta_2 = \frac{\pi}{2}$. Then from [link], the magnetic field along the central axis of an infinite solenoid is

Equation:

$$\vec{\mathbf{B}} = \frac{\mu_0 IN}{2L}\hat{\mathbf{j}}\left[\sin(\pi/2) - \sin(-\pi/2)\right] = \frac{\mu_0 IN}{L}\hat{\mathbf{j}}$$

or

Equation:

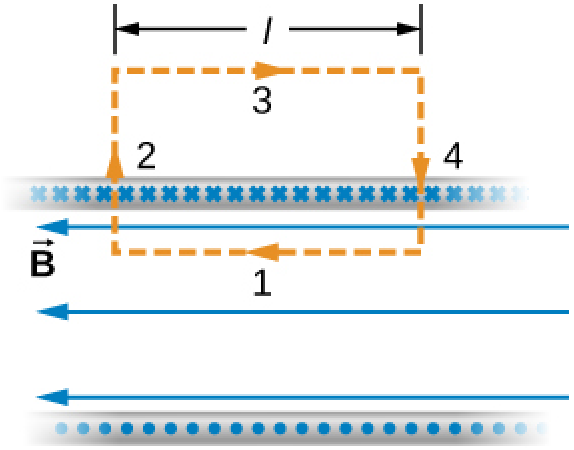
$$\vec{\mathbf{B}} = \mu_0 nI\hat{\mathbf{j}},$$

where *n* is the number of turns per unit length. You can find the direction of $\vec{\mathbf{B}}$ with a right-hand rule: Curl your fingers in the direction of the current, and your thumb points along the magnetic field in the interior of the solenoid.

We now use these properties, along with Ampère’s law, to calculate the magnitude of the magnetic field at any location inside the infinite solenoid. Consider the closed path of [link]. Along segment 1, $\vec{\mathbf{B}}$ is uniform and parallel to the path. Along segments 2 and 4, $\vec{\mathbf{B}}$ is perpendicular to part of the path and vanishes over the rest of it. Therefore, segments 2 and 4 do not contribute to the line integral in Ampère’s law. Along segment 3, $\vec{\mathbf{B}}=0$ because the magnetic field is zero outside the solenoid. If you consider an Ampère’s law loop outside of the solenoid, the current flows in opposite directions on different segments of wire. Therefore, there is no enclosed current and no magnetic field according to Ampère’s law. Thus, there is no contribution to the line integral from segment 3. As a result, we find

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \int_1 \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = Bl.$$



The path of integration used in Ampère’s law to evaluate the magnetic field of an infinite solenoid.

The solenoid has *n* turns per unit length, so the current that passes through the surface enclosed by the path is nlI. Therefore, from Ampère’s law,
Equation:

$$Bl = \mu_0 nlI$$

and

Note:
Equation:

$$B = \mu_0 nI$$

within the solenoid. This agrees with what we found earlier for *B* on the central axis of the solenoid. Here, however, the location of segment 1 is arbitrary, so we have found that this equation gives the magnetic field everywhere inside the infinite solenoid.

Outside the solenoid, one can draw an Ampère’s law loop around the entire solenoid. This would enclose current flowing in both directions. Therefore, the net current inside the loop is zero. According to Ampère’s law, if the net current is zero, the magnetic field must be zero. Therefore, for locations outside of the solenoid’s radius, the magnetic field is zero.

When a patient undergoes a magnetic resonance imaging (MRI) scan, the person lies down on a table that is moved into the center of a large solenoid that can generate very large magnetic fields. The solenoid is capable of these high fields from high currents flowing through superconducting wires. The large magnetic field is used to change the spin of protons in the patient’s body. The time it takes for the spins to align or relax (return to original orientation) is a signature of different tissues that can be analyzed to see if the structures of the tissues is normal ([link]).

In an MRI machine, a large magnetic field is generated by the cylindrical solenoid surrounding the patient. (credit: Liz West)

Example:

Magnetic Field Inside a Solenoid

A solenoid has 300 turns wound around a cylinder of diameter 1.20 cm and length 14.0 cm. If the current through the coils is 0.410 A, what is the magnitude of the magnetic field inside and near the middle of the solenoid?

Strategy

We are given the number of turns and the length of the solenoid so we can find the number of turns per unit length. Therefore, the magnetic field inside and near the middle of the solenoid is given by [link]. Outside the solenoid, the magnetic field is zero.

Solution
The number of turns per unit length is
Equation:

$$n = \frac{300\text{ turns}}{0.140\text{ m}} = 2.14\ \times\ 10^3\text{turns/m.}$$

The magnetic field produced inside the solenoid is
Equation:

$$\begin{aligned} B &= \mu_0 nI = (4\pi\ \times\ 10^{-7}\text{T}\cdot\text{m/A})(2.14\ \times\ 10^3\text{turns/m})(0.410\text{ A}) \\ B &= 1.10\ \times\ 10^{-3}\text{T.} \end{aligned}$$

Significance
This solution is valid only if the length of the solenoid is reasonably large compared with its diameter. This example is a case where this is valid.

Note:
Exercise:

Problem:

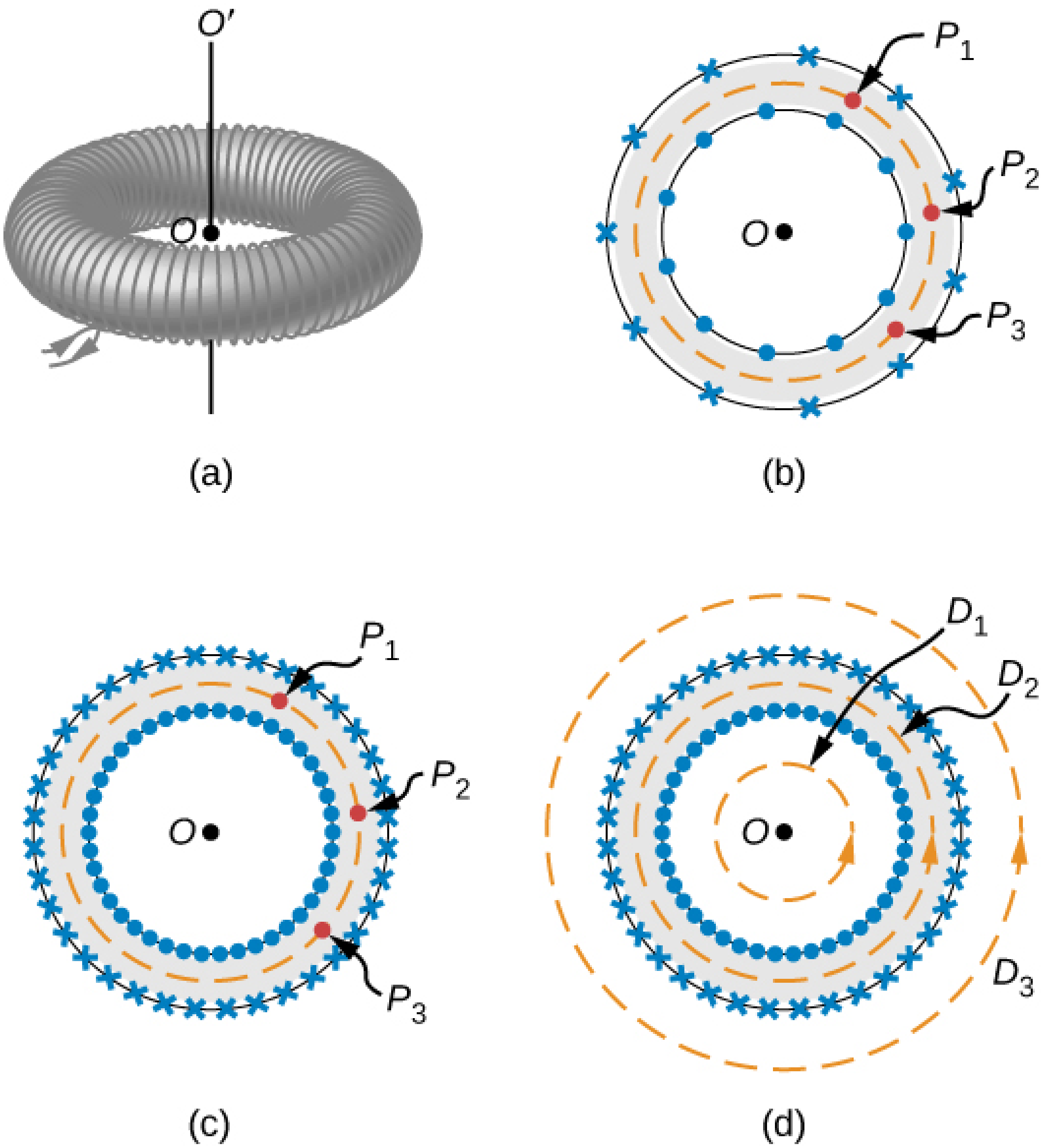
Check Your Understanding What is the ratio of the magnetic field produced from using a finite formula over the infinite approximation for an angle θ of (a) 85°? (b) 89°? The solenoid has 1000 turns in 50 cm with a current of 1.0 A flowing through the coils

Solution:

a. 1.00382; b. 1.00015

Toroids

A toroid is a donut-shaped coil closely wound with one continuous wire, as illustrated in part (a) of [link]. If the toroid has N windings and the current in the wire is I, what is the magnetic field both inside and outside the toroid?



(a) A toroid is a coil wound into a donut-shaped object. (b) A loosely wound toroid does not have cylindrical symmetry. (c) In a tightly wound toroid, cylindrical symmetry is a very good approximation. (d) Several paths of integration for Ampère’s law.

We begin by assuming cylindrical symmetry around the axis *OO*’. Actually, this assumption is not precisely correct, for as part (b) of [link] shows, the view of the toroidal coil varies from point to point (for example, P_1, P_2, and P_3) on a circular path centered around *OO*’. However, if the toroid is tightly wound, all points on the circle become essentially equivalent [part (c) of [link]], and cylindrical symmetry is an accurate approximation.

With this symmetry, the magnetic field must be tangent to and constant in magnitude along any circular path centered on *OO*’. This allows us to write for each of the paths D_1, D_2, and D_3 shown in part (d) of [link],

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = B(2\pi r).$$

Ampère’s law relates this integral to the net current passing through any surface bounded by the path of integration. For a path that is external to the toroid, either no current passes through the enclosing surface (path D_1), or the current passing through the surface in one direction is exactly balanced by the current passing through it in the opposite direction $(\text{path } D_3)$. In either case, there is no net current passing through the surface, so

Equation:

$$\oint B(2\pi r) = 0$$

and

Equation:

$$B = 0 \quad (\text{outside the toroid}).$$

The turns of a toroid form a helix, rather than circular loops. As a result, there is a small field external to the coil; however, the derivation above holds if the coils were circular.

For a circular path within the toroid (path D_2), the current in the wire cuts the surface *N* times, resulting in a net current *NI* through the surface. We now find with Ampère’s law,

Equation:

$$B(2\pi r) = \mu_0 NI$$

and

Note:

Equation:

$$B = \frac{\mu_0 NI}{2\pi r} \quad \text{(within the toroid).}$$

The magnetic field is directed in the counterclockwise direction for the windings shown. When the current in the coils is reversed, the direction of the magnetic field also reverses.

The magnetic field inside a toroid is not uniform, as it varies inversely with the distance *r* from the axis *OO*’. However, if the central radius *R* (the radius midway between the inner and outer radii of the toroid) is much larger than the cross-sectional diameter of the coils *r*, the variation is fairly small, and the magnitude of the magnetic field may be calculated by [link] where $r = R$.

Summary

- The magnetic field strength inside a solenoid is
  Equation:

  $$B = \mu_0 n I \quad \text{(inside a solenoid)}$$

  where n is the number of loops per unit length of the solenoid. The field inside is very uniform in magnitude and direction.
- The magnetic field strength inside a toroid is
  Equation:

  $$B = \frac{\mu_o N I}{2\pi r} \quad \text{(within the toroid)}$$

  where N is the number of windings. The field inside a toroid is not uniform and varies with the distance as $1/r$.

Conceptual Questions

Exercise:

Problem:

Is the magnetic field inside a toroid completely uniform? Almost uniform?

Exercise:

Problem:

Explain why $\vec{\mathbf{B}} = 0$ inside a long, hollow copper pipe that is carrying an electric current parallel to the axis. Is $\vec{\mathbf{B}} = 0$ outside the pipe?

Solution:

If there is no current inside the loop, there is no magnetic field (see Ampère’s law). Outside the pipe, there may be an enclosed current through the copper pipe, so the magnetic field may not be zero outside the pipe.

Problems

Exercise:

Problem:

A solenoid is wound with 2000 turns per meter. When the current is 5.2 A, what is the magnetic field within the solenoid?

Solution:

$B = 1.3 \times 10^{-2}\mathrm{T}$

Exercise:

Problem:

A solenoid has 12 turns per centimeter. What current will produce a magnetic field of $2.0 \times 10^{-2}\mathrm{T}$ within the solenoid?

Exercise:

Problem:

If a current is 2.0 A, how many turns per centimeter must be wound on a solenoid in order to produce a magnetic field of $2.0 \times 10^{-3}\mathrm{T}$ within it?

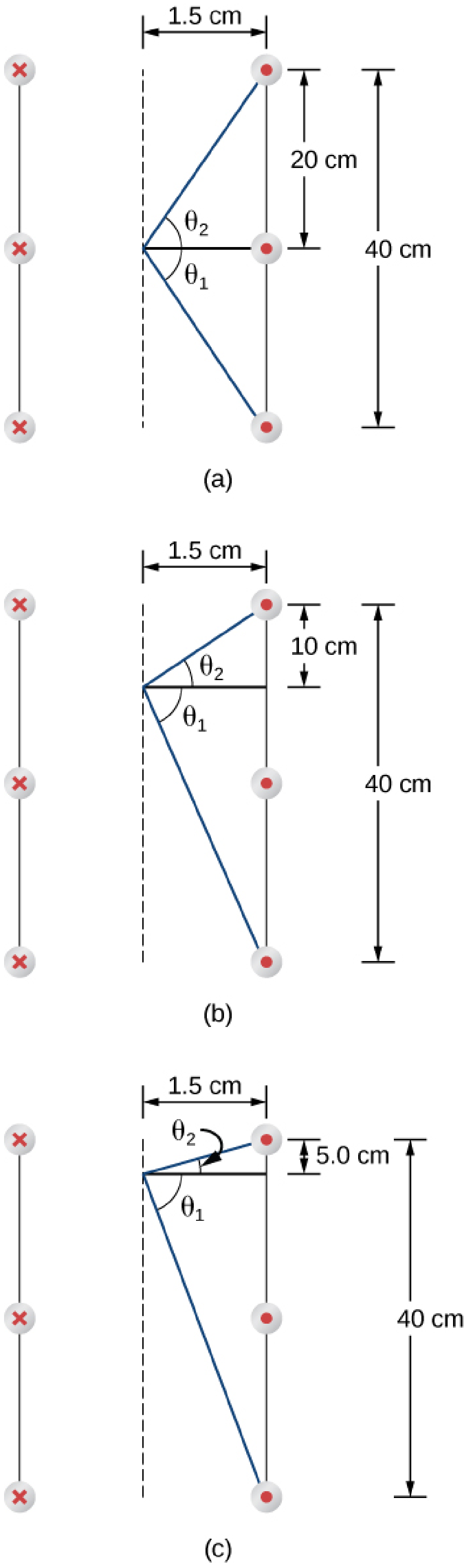
Solution:

roughly eight turns per cm

Exercise:

Problem:

A solenoid is 40 cm long, has a diameter of 3.0 cm, and is wound with 500 turns. If the current through the windings is 4.0 A, what is the magnetic field at a point on the axis of the solenoid that is (a) at the center of the solenoid, (b) 10.0 cm from one end of the solenoid, and (c) 5.0 cm from one end of the solenoid? (d) Compare these answers with the infinite-solenoid case.


1.5 cm
20 cm
40 cm
θ2
θ1
(a)
1.5 cm
10 cm
40 cm
θ2
θ1
(b)
1.5 cm
θ2
5.0 cm
40 cm
θ1
(c)

Exercise:

Problem:

Determine the magnetic field on the central axis at the opening of a semi-infinite solenoid. (That is, take the opening to be at $x = 0$ and the other end to be at
$x = \infty$.)

Solution:

$B = \frac{1}{2}\mu_0 nI$

Exercise:

Problem:

By how much is the approximation $B = \mu_0 nI$ in error at the center of a solenoid that is 15.0 cm long, has a diameter of 4.0 cm, is wrapped with *n* turns per meter, and carries a current *I*?

Exercise:

Problem:

A solenoid with 25 turns per centimeter carries a current *I*. An electron moves within the solenoid in a circle that has a radius of 2.0 cm and is perpendicular to the axis of the solenoid. If the speed of the electron is $2.0 \times 10^5 \text{m/s}$, what is *I*?

Solution:

0.0181 A

Exercise:

Problem:

A toroid has 250 turns of wire and carries a current of 20 A. Its inner and outer radii are 8.0 and 9.0 cm. What are the values of its magnetic field at $r = 8.1, 8.5,$ and 8.9 cm?

Exercise:

Problem:

A toroid with a square cross section 3.0 cm $\times$ 3.0 cm has an inner radius of 25.0 cm. It is wound with 500 turns of wire, and it carries a current of 2.0 A. What is the strength of the magnetic field at the center of the square cross section?

Solution:

0.0008 T

Glossary

solenoid

thin wire wound into a coil that produces a magnetic field when an electric current is passed through it

toroid

donut-shaped coil closely wound around that is one continuous wire

Magnetism in Matter

By the end of this section, you will be able to:

- Classify magnetic materials as paramagnetic, diamagnetic, or ferromagnetic, based on their response to a magnetic field
- Sketch how magnetic dipoles align with the magnetic field in each type of substance
- Define hysteresis and magnetic susceptibility, which determines the type of magnetic material

Why are certain materials magnetic and others not? And why do certain substances become magnetized by a field, whereas others are unaffected? To answer such questions, we need an understanding of magnetism on a microscopic level.

Within an atom, every electron travels in an orbit and spins on an internal axis. Both types of motion produce current loops and therefore magnetic dipoles. For a particular atom, the net magnetic dipole moment is the vector sum of the magnetic dipole moments. Values of μ for several types of atoms are given in [link]. Notice that some atoms have a zero net dipole moment and that the magnitudes of the nonvanishing moments are typically $10^{-23}\text{A}\cdot\text{m}^2$.

Atom	**Magnetic Moment** $\left(10^{-24}\ \text{A}\cdot\text{m}^2\right)$
H	9.27
He	0
Li	9.27
O	13.9

Atom	**Magnetic Moment** $\left(10^{-24}\ \text{A}\cdot\text{m}^2\right)$
Na	9.27
S	13.9

Magnetic Moments of Some Atoms

A handful of matter has approximately 10^{26} atoms and ions, each with its magnetic dipole moment. If no external magnetic field is present, the magnetic dipoles are randomly oriented—as many are pointed up as down, as many are pointed east as west, and so on. Consequently, the net magnetic dipole moment of the sample is zero. However, if the sample is placed in a magnetic field, these dipoles tend to align with the field (see [link]), and this alignment determines how the sample responds to the field. On the basis of this response, a material is said to be either paramagnetic, ferromagnetic, or diamagnetic.

In a **paramagnetic material**, only a small fraction (roughly one-third) of the magnetic dipoles are aligned with the applied field. Since each dipole produces its own magnetic field, this alignment contributes an extra magnetic field, which enhances the applied field. When a **ferromagnetic material** is placed in a magnetic field, its magnetic dipoles also become aligned; furthermore, they become locked together so that a permanent magnetization results, even when the field is turned off or reversed. This permanent magnetization happens in ferromagnetic materials but not paramagnetic materials. **Diamagnetic materials** are composed of atoms that have no net magnetic dipole moment. However, when a diamagnetic material is placed in a magnetic field, a magnetic dipole moment is directed opposite to the applied field and therefore produces a magnetic field that opposes the applied field. We now consider each type of material in greater detail.

Paramagnetic Materials

For simplicity, we assume our sample is a long, cylindrical piece that completely fills the interior of a long, tightly wound solenoid. When there is no current in the solenoid, the magnetic dipoles in the sample are randomly oriented and produce no net magnetic field. With a solenoid current, the magnetic field due to the solenoid exerts a torque on the dipoles that tends to align them with the field. In competition with the aligning torque are thermal collisions that tend to randomize the orientations of the dipoles. The relative importance of these two competing processes can be estimated by comparing the energies involved. From [link], the energy difference between a magnetic dipole aligned with and against a magnetic field is $U_B = 2\mu B$. If $\mu = 9.3 \times 10^{-24} \text{A} \cdot \text{m}^2$ (the value of atomic hydrogen) and B = 1.0 T, then

Equation:

$$U_B = 1.9 \times 10^{-23} \text{J}.$$

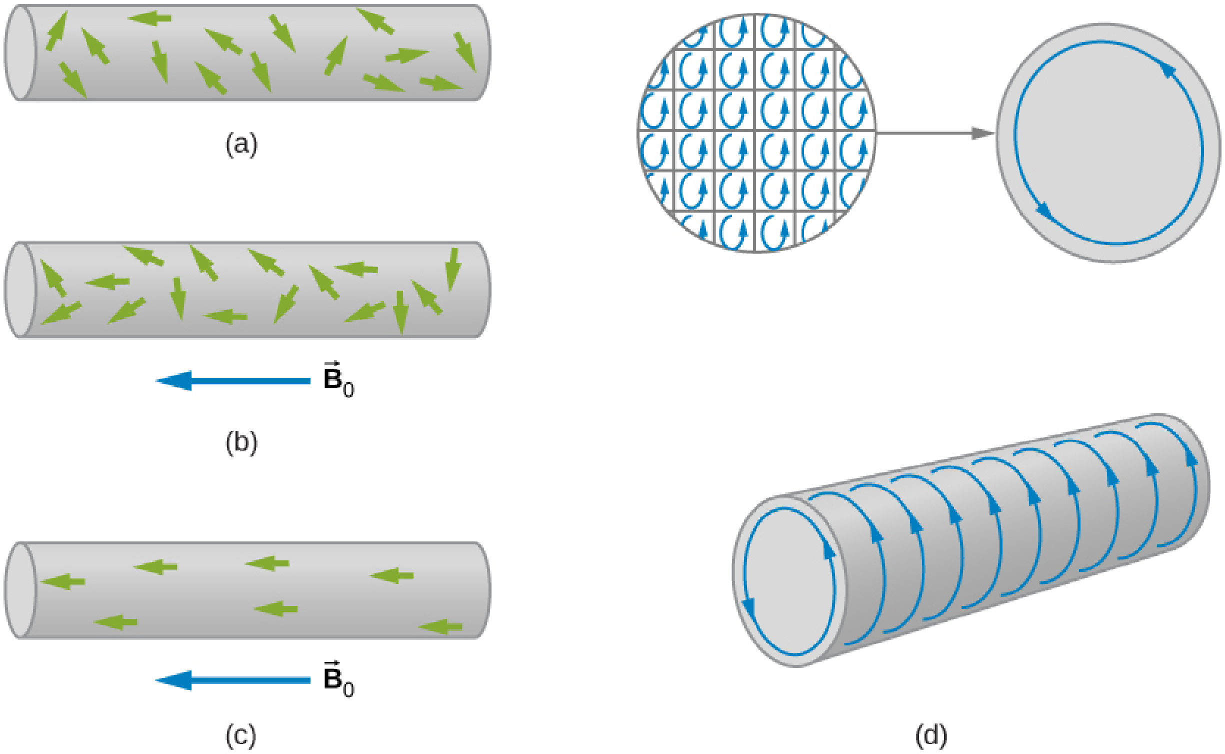
At a room temperature of $27\ ^\circ\text{C}$, the thermal energy per atom is

Equation:

$$U_T \approx kT = (1.38 \times 10^{-23} \text{J/K})(300\,\text{K}) = 4.1 \times 10^{-21} \text{J},$$

which is about 220 times greater than U_B. Clearly, energy exchanges in thermal collisions can seriously interfere with the alignment of the magnetic dipoles. As a result, only a small fraction of the dipoles is aligned at any instant.

The four sketches of [link] furnish a simple model of this alignment process. In part (a), before the field of the solenoid (not shown) containing the paramagnetic sample is applied, the magnetic dipoles are randomly oriented and there is no net magnetic dipole moment associated with the material. With the introduction of the field, a partial alignment of the dipoles takes place, as depicted in part (b). The component of the net magnetic dipole moment that is perpendicular to the field vanishes. We may then represent the sample by part (c), which shows a collection of magnetic dipoles completely aligned with the field. By treating these dipoles as current loops, we can picture the dipole alignment as equivalent to a current around the surface of the material, as in part (d). This fictitious surface current produces its own magnetic field, which enhances the field of the solenoid.



The alignment process in a paramagnetic material filling a solenoid (not shown). (a) Without an applied field, the magnetic dipoles are randomly oriented. (b) With a field, partial alignment occurs. (c) An equivalent representation of part (b). (d) The internal currents cancel, leaving an effective surface current that produces a magnetic field similar to that of a finite solenoid.

We can express the total magnetic field $\vec{\mathbf{B}}$ in the material as

Equation:

$$\vec{\mathbf{B}} = \vec{\mathbf{B}}_0 + \vec{\mathbf{B}}_m,$$

where $\vec{\mathbf{B}}_0$ is the field due to the current I_0 in the solenoid and $\vec{\mathbf{B}}_m$ is the field due to the surface current I_m around the sample. Now $\vec{\mathbf{B}}_m$ is usually proportional to $\vec{\mathbf{B}}_0$, a fact we express by

Equation:

$$\vec{\mathbf{B}}_m = \chi \vec{\mathbf{B}}_0,$$

where χ is a dimensionless quantity called the **magnetic susceptibility**. Values of χ for some paramagnetic materials are given in [link]. Since the alignment of magnetic dipoles is so weak, χ is very small for paramagnetic materials. By combining [link] and [link], we obtain:

Equation:

$$\vec{\mathbf{B}} = \vec{\mathbf{B}}_0 + \chi \vec{\mathbf{B}}_0 = (1 + \chi)\vec{\mathbf{B}}_0.$$

For a sample within an infinite solenoid, this becomes

Equation:

$$B = (1 + \chi)\mu_0 n I.$$

This expression tells us that the insertion of a paramagnetic material into a solenoid increases the field by a factor of $(1 + \chi)$. However, since χ is so small, the field isn't enhanced very much.

The quantity

Note:

Equation:

$$\mu = (1 + \chi)\mu_0.$$

is called the magnetic permeability of a material. In terms of μ, [link] can be written as

Note:
Equation:

$$B = \mu n I$$

for the filled solenoid.

Paramagnetic Materials	χ	**Diamagnetic Materials**	χ
Aluminum	2.2×10^{-5}	Bismuth	-1.7×10^{-5}
Calcium	1.4×10^{-5}	Carbon (diamond)	-2.2×10^{-5}
Chromium	3.1×10^{-4}	Copper	-9.7×10^{-6}
Magnesium	1.2×10^{-5}	Lead	-1.8×10^{-5}
Oxygen gas (1 atm)	1.8×10^{-6}	Mercury	-2.8×10^{-5}
Oxygen liquid (90 K)	3.5×10^{-3}	Hydrogen gas (1 atm)	-2.2×10^{-9}

Paramagnetic Materials	χ	**Diamagnetic Materials**	χ
Tungsten	$6.8\,\times\,10^{-5}$	Nitrogen gas (1 atm)	$-6.7\,\times\,10^{-9}$
Air (1 atm)	$3.6\,\times\,10^{-7}$	Water	$-9.1\,\times\,10^{-6}$

Magnetic Susceptibilities*Note: Unless otherwise specified, values given are for room temperature.

Diamagnetic Materials

A magnetic field always induces a magnetic dipole in an atom. This induced dipole points opposite to the applied field, so its magnetic field is also directed opposite to the applied field. In paramagnetic and ferromagnetic materials, the induced magnetic dipole is masked by much stronger permanent magnetic dipoles of the atoms. However, in diamagnetic materials, whose atoms have no permanent magnetic dipole moments, the effect of the induced dipole is observable.

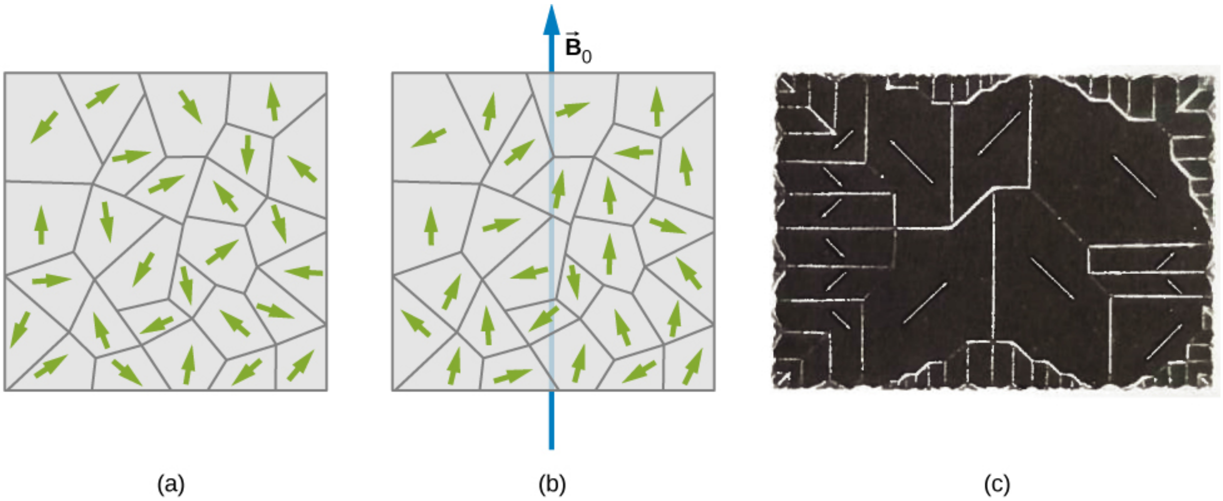
We can now describe the magnetic effects of diamagnetic materials with the same model developed for paramagnetic materials. In this case, however, the fictitious surface current flows opposite to the solenoid current, and the magnetic susceptibility χ is negative. Values of χ for some diamagnetic materials are also given in [link].

Note:
Water is a common diamagnetic material. Animals are mostly composed of water. Experiments have been performed on frogs and mice in diverging magnetic fields. The water molecules are repelled from the applied magnetic field against gravity until the animal reaches an equilibrium. The result is that the animal is levitated by the magnetic field.

Ferromagnetic Materials

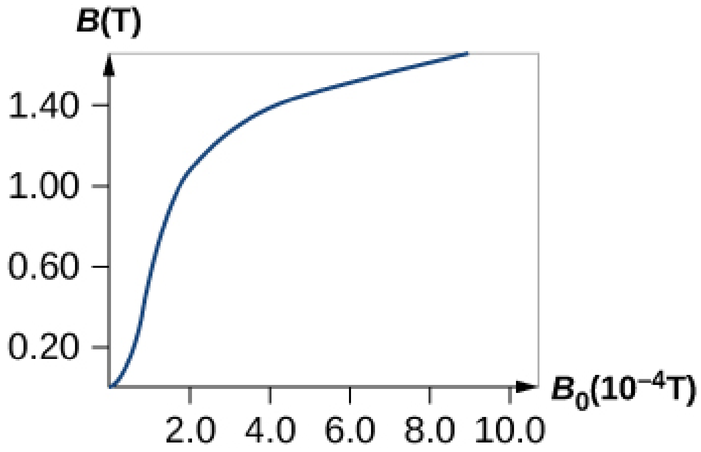
Common magnets are made of a ferromagnetic material such as iron or one of its alloys. Experiments reveal that a ferromagnetic material consists of tiny regions known as **magnetic domains**. Their volumes typically range from 10^{-12} to 10^{-8}m^3, and they contain about 10^{17} to 10^{21} atoms. Within a domain, the magnetic dipoles are rigidly aligned in the same direction by coupling among the atoms. This coupling, which is due to quantum mechanical effects, is so strong that even thermal agitation at room temperature cannot break it. The result is that each domain has a net dipole moment. Some materials have weaker coupling and are ferromagnetic only at lower temperatures.

If the domains in a ferromagnetic sample are randomly oriented, as shown in [link], the sample has no net magnetic dipole moment and is said to be unmagnetized. Suppose that we fill the volume of a solenoid with an unmagnetized ferromagnetic sample. When the magnetic field $\vec{\mathbf{B}}_0$ of the solenoid is turned on, the dipole moments of the domains rotate so that they align somewhat with the field, as depicted in [link]. In addition, the aligned domains tend to increase in size at the expense of unaligned ones. The net effect of these two processes is the creation of a net magnetic dipole moment for the ferromagnet that is directed along the applied magnetic field. This net magnetic dipole moment is much larger than that of a paramagnetic sample, and the domains, with their large numbers of atoms, do not become misaligned by thermal agitation. Consequently, the field due to the alignment of the domains is quite large.



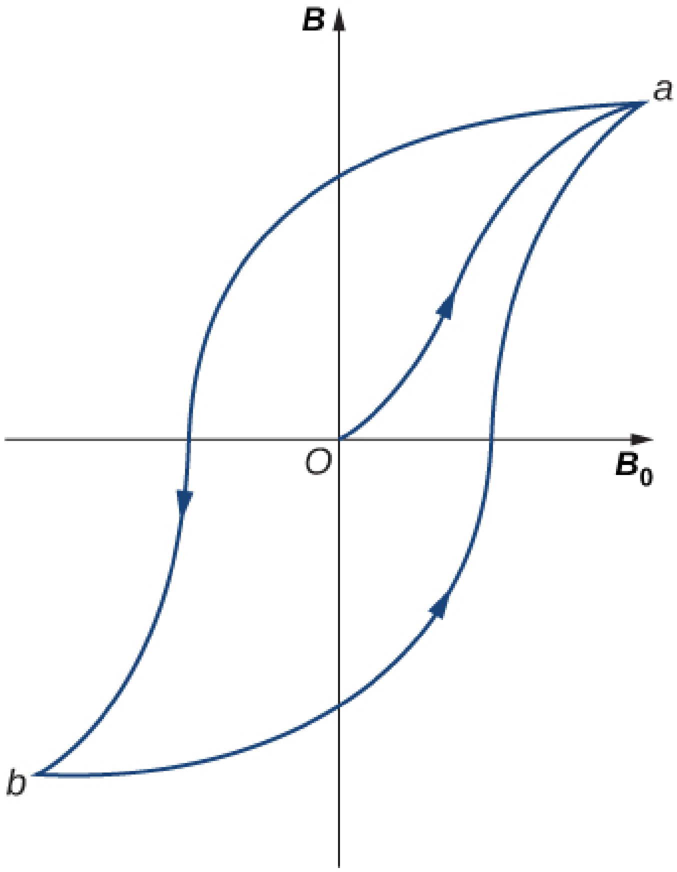
(a) Domains are randomly oriented in an unmagnetized ferromagnetic sample such as iron. The arrows represent the orientations of the magnetic dipoles within the domains. (b) In an applied magnetic field, the domains align somewhat with the field. (c) The domains of a single crystal of nickel. The white lines show the boundaries of the domains. These lines are produced by iron oxide powder sprinkled on the crystal.

Besides iron, only four elements contain the magnetic domains needed to exhibit ferromagnetic behavior: cobalt, nickel, gadolinium, and dysprosium. Many alloys of these elements are also ferromagnetic. Ferromagnetic materials can be described using [link] through [link], the paramagnetic equations. However, the value of χ for ferromagnetic material is usually on the order of 10^3 to 10^4, and it also depends on the history of the magnetic field to which the material has been subject. A typical plot of B (the total field in the material) versus B_0 (the applied field) for an initially unmagnetized piece of iron is shown in [link]. Some sample numbers are (1) for $B_0 = 1.0 \times 10^{-4} \text{T}, B = 0.60 \text{ T}$, and $\chi = \left({}^{0.60}/_{1.0 \times 10^{-4}}\right) - 1 \approx 6.0 \times 10^3$; (2) for $B_0 = 6.0 \times 10^{-4} \text{T}, B = 1.5 \text{ T}$, and $\chi = \left({}^{1.5}/_{6.0 \times 10^{-4}}\right) - 1 \approx 2.5 \times 10^3$.



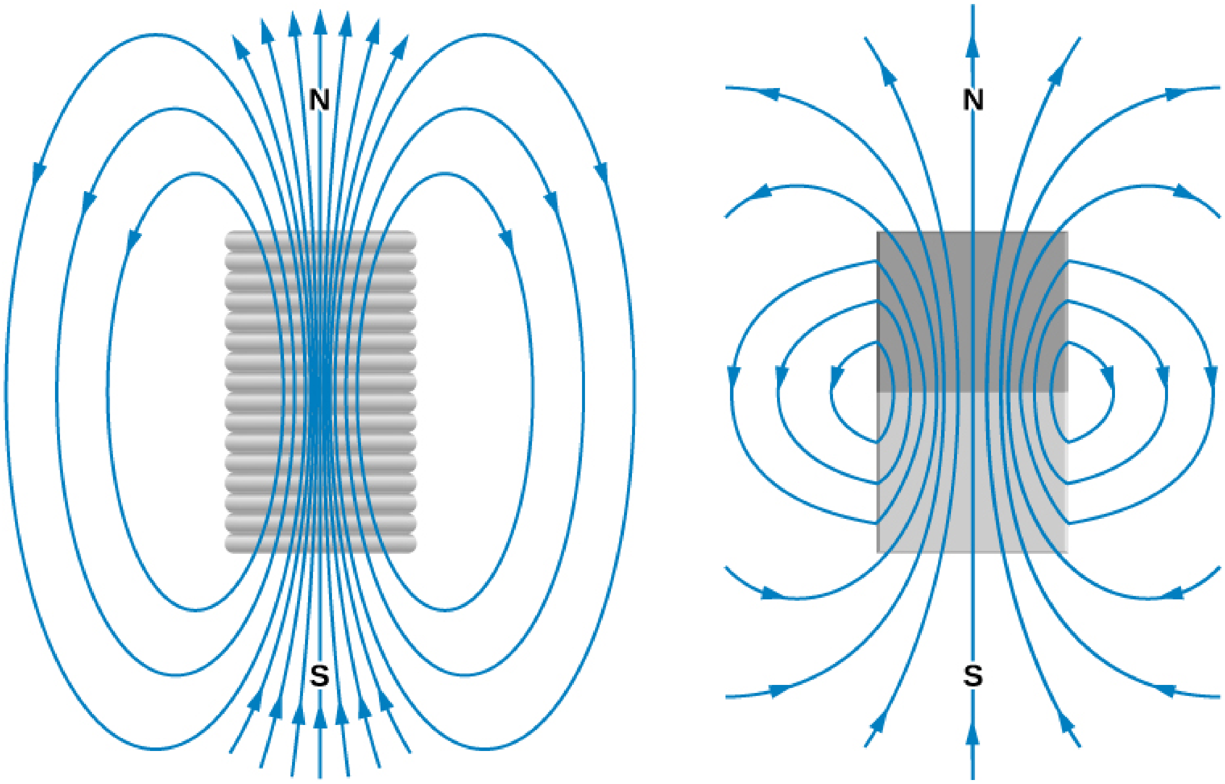
(a) The magnetic field *B* in annealed iron as a function of the applied field B_0.

When B_0 is varied over a range of positive and negative values, *B* is found to behave as shown in [link]. Note that the same B_0 (corresponding to the same current in the solenoid) can produce different values of *B* in the material. The magnetic field *B* produced in a ferromagnetic material by an applied field B_0 depends on the magnetic history of the material. This effect is called **hysteresis**, and the curve of [link] is called a hysteresis loop. Notice that *B* does not disappear when $B_0 = 0$ (i.e., when the current in the solenoid is turned off). The iron stays magnetized, which means that it has become a permanent magnet.



A typical hysteresis loop for a ferromagnet. When the material is first magnetized, it follows a curve from 0 to *a*. When B_0 is reversed, it takes the path shown from *a* to *b*. If B_0 is reversed again, the material follows the curve from *b* to *a*.

Like the paramagnetic sample of [link], the partial alignment of the domains in a ferromagnet is equivalent to a current flowing around the surface. A bar magnet can therefore be pictured as a tightly wound solenoid with a large current circulating through its coils (the surface current). You can see in [link] that this model fits quite well. The fields of the bar magnet and the finite solenoid are strikingly similar. The figure also shows how the poles of the bar magnet are identified. To form closed loops, the field lines outside the magnet leave the north (N) pole and enter the south (S) pole, whereas inside the magnet, they leave S and enter N.



Comparison of the magnetic fields of a finite solenoid and a bar magnet.

Ferromagnetic materials are found in computer hard disk drives and permanent data storage devices ([link]). A material used in your hard disk drives is called a spin valve, which has alternating layers of ferromagnetic (aligning with the external magnetic field) and antiferromagnetic (each atom is aligned opposite to the next) metals. It was observed that a significant change in resistance was discovered based on whether an applied magnetic field was on the spin valve or not. This large change in resistance creates a quick and consistent way for recording or reading information by an applied current.

The inside of a hard disk drive. The silver disk contains the information, whereas the thin stylus on top of the disk reads and writes information to the disk.

Example:

Iron Core in a Coil

A long coil is tightly wound around an iron cylinder whose magnetization curve is shown in [link]. (a) If $n = 20$ turns per centimeter, what is the applied field B_0 when $I_0 = 0.20$ A? (b) What is the net magnetic field for this same current? (c) What is the magnetic susceptibility in this case?

Strategy

(a) The magnetic field of a solenoid is calculated using [link]. (b) The graph is read to determine the net magnetic field for this same current. (c) The magnetic susceptibility is calculated using [link].

Solution

a. The applied field B_0 of the coil is

Equation:

$$\begin{aligned} B_0 &= \mu_0 n I_0 = (4\pi \times 10^{-7} \mathrm{T \cdot m/A})(2000\,/\,\mathrm{m})(0.20\ \mathrm{A}) \\ B_0 &= 5.0 \times 10^{-4} \mathrm{T}. \end{aligned}$$

b. From inspection of the magnetization curve of [link], we see that, for this value of B_0, $B = 1.4\ \mathrm{T}$. Notice that the internal field of the aligned atoms is much larger than the externally applied field.

c. The magnetic susceptibility is calculated to be

Equation:

$$\chi = \frac{B}{B_0} - 1 = \frac{1.4\ \mathrm{T}}{5.0 \times 10^{-4} \mathrm{T}} - 1 = 2.8 \times 10^3.$$

Significance

Ferromagnetic materials have susceptibilities in the range of 10^3 which compares well to our results here. Paramagnetic materials have fractional susceptibilities, so their applied field of the coil is much greater than the magnetic field generated by the material.

Note:

Exercise:

Problem:

Check Your Understanding Repeat the calculations from the previous example for $I_0 = 0.040\ \mathrm{A}$.

Solution:

a. $1.0 \times 10^{-4} \mathrm{T}$; b. $0.60\ \mathrm{T}$; c. 6.0×10^3

Summary

- Materials are classified as paramagnetic, diamagnetic, or ferromagnetic, depending on how they behave in an applied magnetic field.
- Paramagnetic materials have partial alignment of their magnetic dipoles with an applied magnetic field. This is a positive magnetic susceptibility. Only a surface current remains, creating a solenoid-like magnetic field.
- Diamagnetic materials exhibit induced dipoles opposite to an applied magnetic field. This is a negative magnetic susceptibility.
- Ferromagnetic materials have groups of dipoles, called domains, which align with the applied magnetic field. However, when the field is removed, the ferromagnetic material remains magnetized, unlike paramagnetic materials. This magnetization of the material versus the applied field effect is called hysteresis.

Key Equations

Permeability of free space	$\mu_0 = 4\pi \times 10^{-7} \mathrm{T \cdot m/A}$
Contribution to magnetic field from a current element	$dB = \frac{\mu_0}{4\pi} \frac{I\,dl \sin\theta}{r^2}$
Biot–Savart law	$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \int\limits_{\text{wire}} \frac{Id\vec{\mathbf{l}} \times \hat{\mathbf{r}}}{r^2}$
Magnetic field due to a long straight wire	$B = \frac{\mu_0 I}{2\pi R}$
Force between two parallel	

currents	$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}$
Magnetic field of a current loop	$B = \frac{\mu_0 I}{2R}$ (at center of loop)
Ampère’s law	$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = \mu_0 I$
Magnetic field strength inside a solenoid	$B = \mu_0 n I$
Magnetic field strength inside a toroid	$B = \frac{\mu_o N I}{2\pi r}$
Magnetic permeability	$\mu = (1 + \chi)\mu_0$
Magnetic field of a solenoid filled with paramagnetic material	$B = \mu n I$

Conceptual Questions

Exercise:

Problem:

A diamagnetic material is brought close to a permanent magnet. What happens to the material?

Exercise:

Problem:

If you cut a bar magnet into two pieces, will you end up with one magnet with an isolated north pole and another magnet with an isolated south pole? Explain your answer.

Solution:

The bar magnet will then become two magnets, each with their own north and south poles. There are no magnetic monopoles or single pole magnets.

Problems

Exercise:

Problem:

The magnetic field in the core of an air-filled solenoid is 1.50 T. By how much will this magnetic field decrease if the air is pumped out of the core while the current is held constant?

Exercise:

Problem:

A solenoid has a ferromagnetic core, n = 1000 turns per meter, and I = 5.0 A. If B inside the solenoid is 2.0 T, what is χ for the core material?

Solution:

317.31

Exercise:

Problem:

A 20-A current flows through a solenoid with 2000 turns per meter. What is the magnetic field inside the solenoid if its core is (a) a vacuum and (b) filled with liquid oxygen at 90 K?

Exercise:

Problem:

The magnetic dipole moment of the iron atom is about $2.1 \times 10^{-23} \text{A} \cdot \text{m}^2$. (a) Calculate the maximum magnetic dipole moment of a domain consisting of 10^{19} iron atoms. (b) What current would have to flow through a single circular loop of wire of diameter 1.0 cm to produce this magnetic dipole moment?

Solution:

$2.1 \times 10^{-4} \text{A} \cdot \text{m}^2$
2.7 A

Exercise:

Problem:

Suppose you wish to produce a 1.2-T magnetic field in a toroid with an iron core for which $\chi = 4.0 \times 10^3$. The toroid has a mean radius of 15 cm and is wound with 500 turns. What current is required?

Exercise:

Problem:

A current of 1.5 A flows through the windings of a large, thin toroid with 200 turns per meter and a radius of 1 meter. If the toroid is filled with iron for which $\chi = 3.0 \times 10^3$, what is the magnetic field within it?

Solution:

0.18 T

Exercise:

Problem:

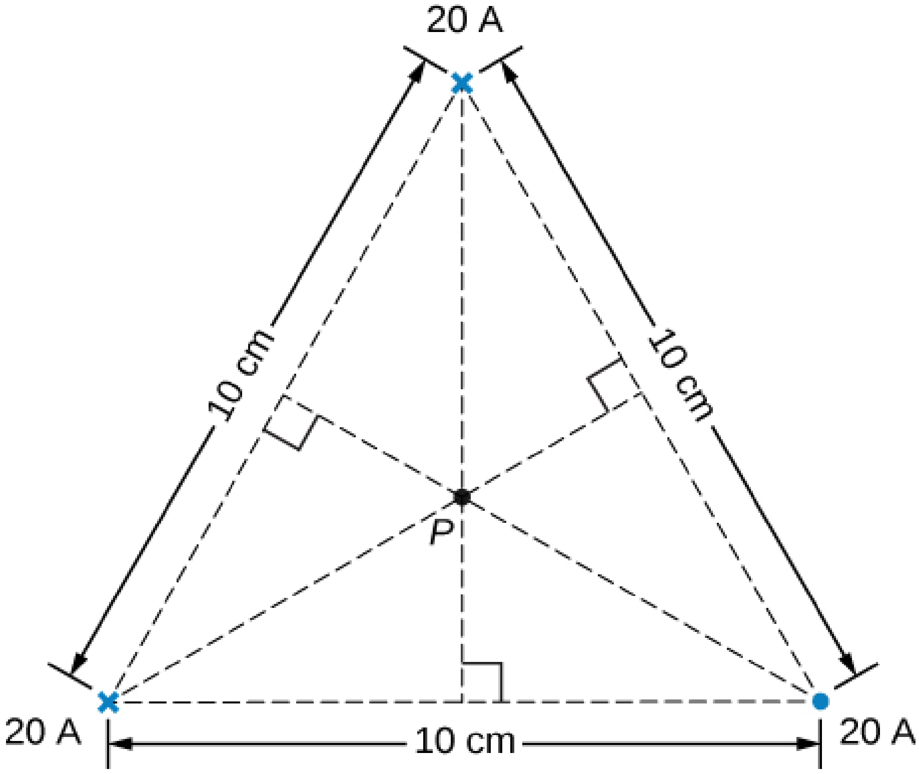
A solenoid with an iron core is 25 cm long and is wrapped with 100 turns of wire. When the current through the solenoid is 10 A, the magnetic field inside it is 2.0 T. For this current, what is the permeability of the iron? If the current is turned off and then restored to 10 A, will the magnetic field necessarily return to 2.0 T?

Additional Problems

Exercise:

Problem:

Three long, straight, parallel wires, all carrying 20 A, are positioned as shown in the accompanying figure. What is the magnitude of the magnetic field at the point *P*?



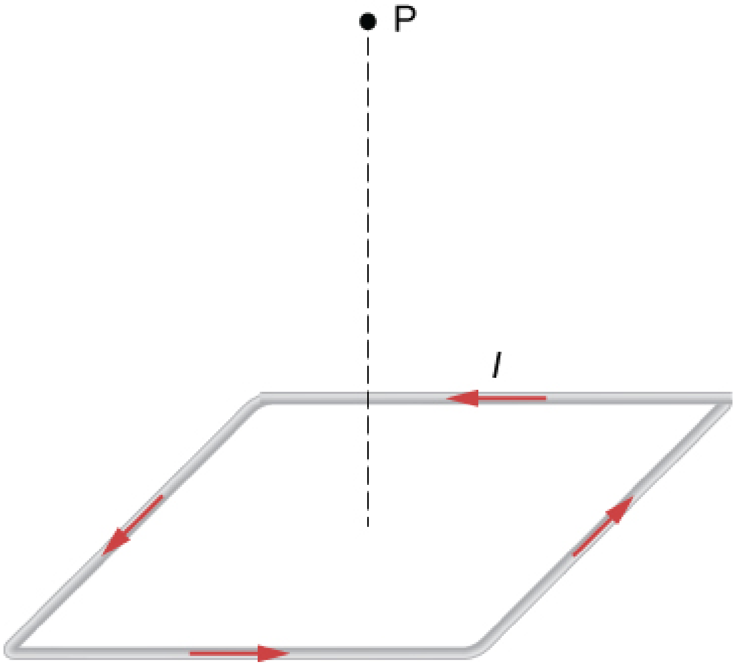
Solution:

$B=1.4\ \times\ 10^{-4}\mathrm{T}$

Exercise:

Problem:

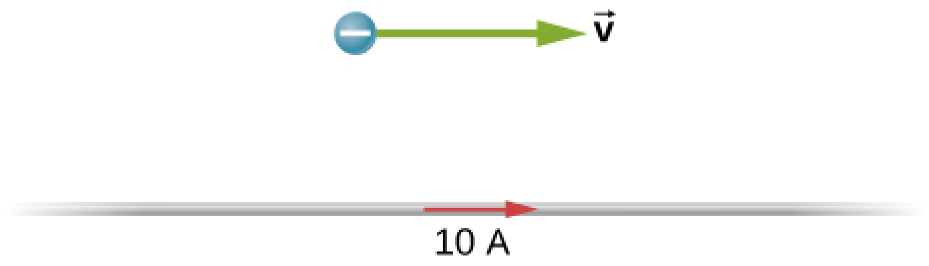
A current I flows around a wire bent into the shape of a square of side a. What is the magnetic field at the point P that is a distance z above the center of the square (see the accompanying figure)?



Exercise:

Problem:

The accompanying figure shows a long, straight wire carrying a current of 10 A. What is the magnetic force on an electron at the instant it is 20 cm from the wire, traveling parallel to the wire with a speed of $2.0\ \times\ 10^{5}\mathrm{m/s}$? Describe qualitatively the subsequent motion of the electron.



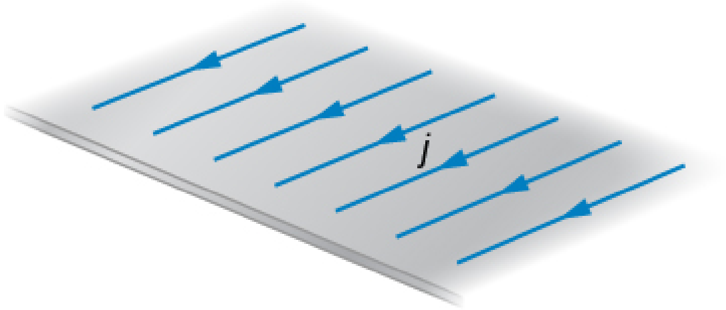
Solution:

$3.2 \times 10^{-19} N$ in an arc away from the wire

Exercise:

Problem:

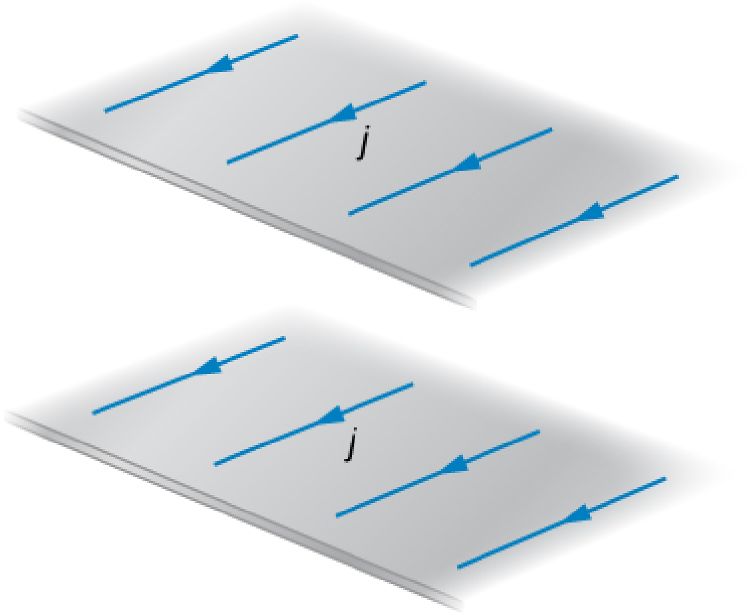
Current flows along a thin, infinite sheet as shown in the accompanying figure. The current per unit length along the sheet is J in amperes per meter. (a) Use the Biot-Savart law to show that $B = \mu_0 J/2$ on either side of the sheet. What is the direction of $\vec{\mathbf{B}}$ on each side? (b) Now use Ampère’s law to calculate the field.



Exercise:

Problem:

(a) Use the result of the previous problem to calculate the magnetic field between, above, and below the pair of infinite sheets shown in the accompanying figure. (b) Repeat your calculations if the direction of the current in the lower sheet is reversed.



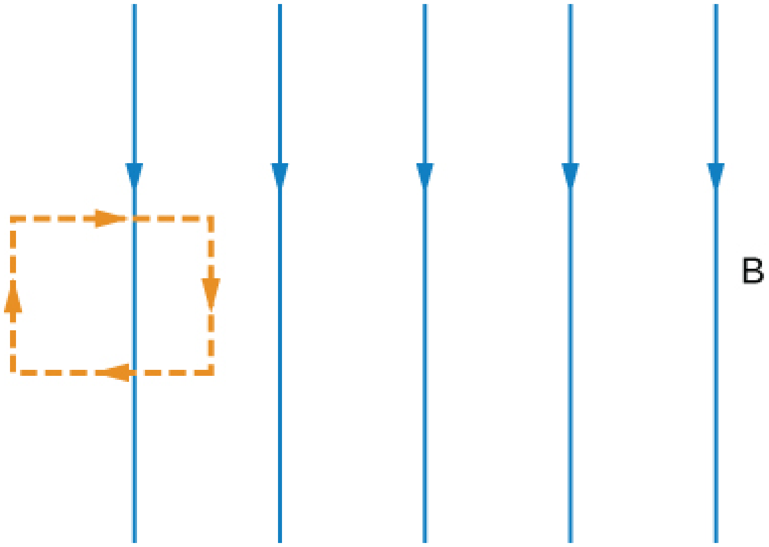
Solution:

a. above and below $B = \mu_0 j$, in the middle $B = 0$; b. above and below $B = 0$, in the middle $B = \mu_0 j$

Exercise:

Problem:

We often assume that the magnetic field is uniform in a region and zero everywhere else. Show that in reality it is impossible for a magnetic field to drop abruptly to zero, as illustrated in the accompanying figure. (*Hint*: Apply Ampère’s law over the path shown.)



Exercise:

Problem:

How is the fractional change in the strength of the magnetic field across the face of the toroid related to the fractional change in the radial distance from the axis of the toroid?

Solution:

$\frac{dB}{B} = -\frac{dr}{r}$

Exercise:

Problem:

Show that the expression for the magnetic field of a toroid reduces to that for the field of an infinite solenoid in the limit that the central radius goes to infinity.

Exercise:

Problem:

A toroid with an inner radius of 20 cm and an outer radius of 22 cm is tightly wound with one layer of wire that has a diameter of 0.25 mm. (a) How many turns are there on the toroid? (b) If the current through the toroid windings is 2.0 A, what is the strength of the magnetic field at the center of the toroid?

Solution:

a. 5026 turns; b. 0.00957 T

Exercise:

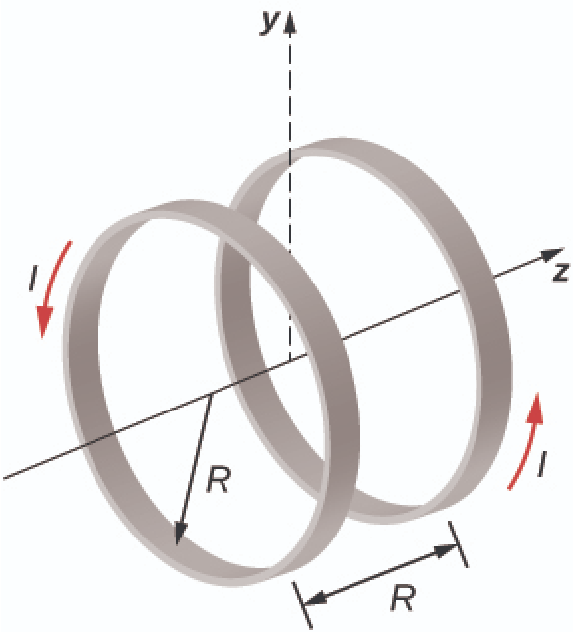
Problem:

A wire element has $d\vec{\mathbf{l}}$, $Id\vec{\mathbf{l}} = \mathbf{J}Adl = \mathbf{J}dv$, where A and dv are the cross-sectional area and volume of the element, respectively. Use this, the Biot-Savart law, and $\mathbf{J} = ne\mathbf{v}$ to show that the magnetic field of a moving point charge q is given by:
$\vec{\mathbf{B}} = \frac{\mu_0}{4\pi} \frac{q\mathbf{v} \times \hat{\mathbf{r}}}{r^2}$

Exercise:

Problem:

A reasonably uniform magnetic field over a limited region of space can be produced with the Helmholtz coil, which consists of two parallel coils centered on the same axis. The coils are connected so that they carry the same current I. Each coil has N turns and radius R, which is also the distance between the coils. (a) Find the magnetic field at any point on the z-axis shown in the accompanying figure. (b) Show that dB/dz and ${}^{d^2B}\!/_{dz^2}$ are both zero at $z = 0$. (These vanishing derivatives demonstrate that the magnetic field varies only slightly near $z = 0$.)



Solution:

$$B_1(x) = \frac{\mu_0 I R^2}{2(R^2+z^2)^{3/2}}$$

Exercise:

Problem:

A charge of 4.0 μC is distributed uniformly around a thin ring of insulating material. The ring has a radius of 0.20 m and rotates at $2.0 \times 10^4 \text{rev/min}$ around the axis that passes through its center and is perpendicular to the plane of the ring. What is the magnetic field at the center of the ring?

Exercise:

Problem:

A thin, nonconducting disk of radius R is free to rotate around the axis that passes through its center and is perpendicular to the face of the disk. The disk is charged uniformly with a total charge q. If the disk rotates at a constant angular velocity ω, what is the magnetic field at its center?

Solution:

$B = \frac{\mu_0 \sigma \omega}{2} R$

Exercise:

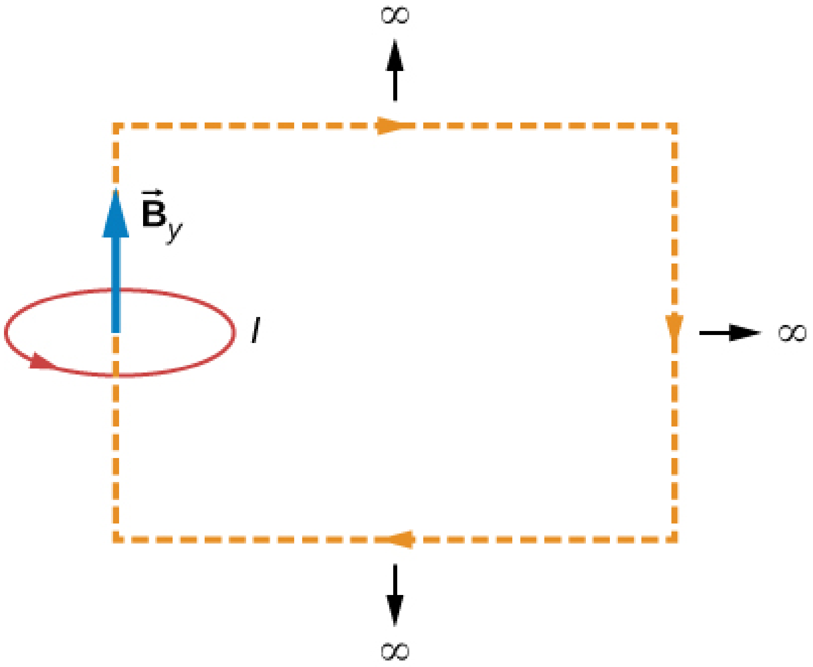
Problem:

Consider the disk in the previous problem. Calculate the magnetic field at a point on its central axis that is a distance *y* above the disk.

Exercise:

Problem:

Consider the axial magnetic field $B_y = \mu_0 I R^2 / 2(y^2 + R^2)^{3/2}$ of the circular current loop shown below. (a) Evaluate $\int_{-a}^{a} B_y dy$. Also show that $\lim_{a \to \infty} \int_{-a}^{a} B_y dy = \mu_0 I$. (b) Can you deduce this limit without evaluating the integral? (*Hint:* See the accompanying figure.)



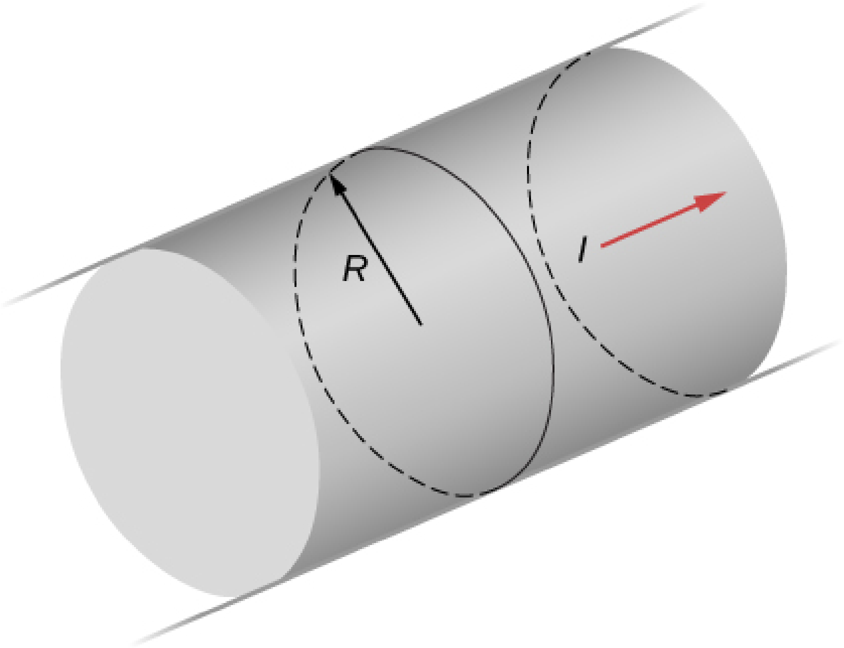
Solution:

derivation

Exercise:

Problem:

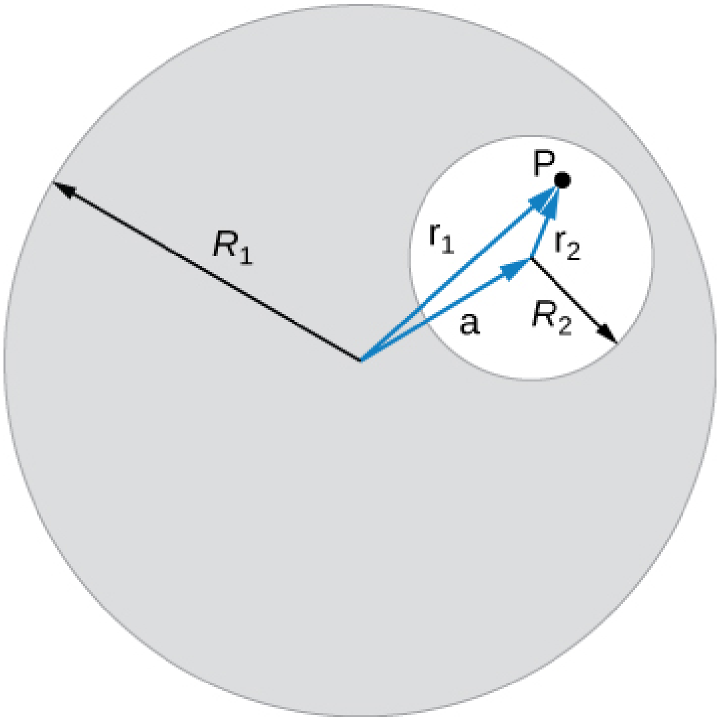
The current density in the long, cylindrical wire shown in the accompanying figure varies with distance r from the center of the wire according to $J = cr$, where c is a constant. (a) What is the current through the wire? (b) What is the magnetic field produced by this current for $r \leq R$? For $r \geq R$?



Exercise:

Problem:

A long, straight, cylindrical conductor contains a cylindrical cavity whose axis is displaced by *a* from the axis of the conductor, as shown in the accompanying figure. The current density in the conductor is given by $\vec{\mathbf{J}} = J_0\widehat{\mathbf{k}}$, where J_0 is a constant and $\widehat{\mathbf{k}}$ is along the axis of the conductor. Calculate the magnetic field at an arbitrary point P in the cavity by superimposing the field of a solid cylindrical conductor with radius R_1 and current density $\vec{\mathbf{J}}$ onto the field of a solid cylindrical conductor with radius R_2 and current density $-\vec{\mathbf{J}}$. Then use the fact that the appropriate azimuthal unit vectors can be expressed as $\hat{\theta}_1 = \widehat{k} \times \hat{r}_1$ and $\hat{\theta}_2 = \widehat{k} \times \hat{r}_2$ to show that everywhere inside the cavity the magnetic field is given by the constant $\vec{\mathbf{B}} = \frac{1}{2}\mu_0 J_0 \mathbf{k} \times \mathbf{a}$, where $\mathbf{a} = \mathbf{r}_1 - \mathbf{r}_2$ and $\mathbf{r}_1 = r_1\hat{r}_1$ is the position of *P* relative to the center of the conductor and $\mathbf{r}_2 = r_2\hat{r}_2$ is the position of *P* relative to the center of the cavity.



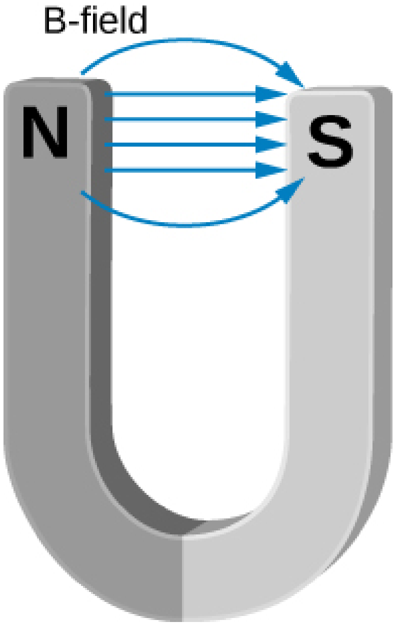
Solution:

derivation

Exercise:

Problem:

Between the two ends of a horseshoe magnet the field is uniform as shown in the diagram. As you move out to outside edges, the field bends. Show by Ampère’s law that the field must bend and thereby the field weakens due to these bends.



Exercise:

Problem:

Show that the magnetic field of a thin wire and that of a current loop are zero if you are infinitely far away.

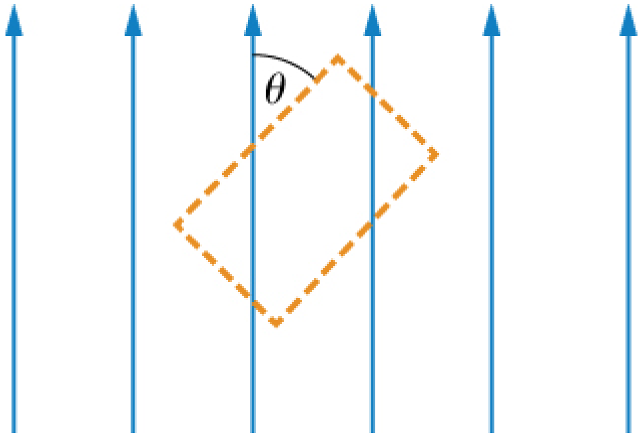
Solution:

As the radial distance goes to infinity, the magnetic fields of each of these formulae go to zero.

Exercise:

Problem:

An Ampère loop is chosen as shown by dashed lines for a parallel constant magnetic field as shown by solid arrows. Calculate $\vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$ for each side of the loop then find the entire $\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}}$. Can you think of an Ampère loop that would make the problem easier? Do those results match these?



Exercise:

Problem:

A very long, thick cylindrical wire of radius R carries a current density J that varies across its cross-section. The magnitude of the current density at a point a distance r from the center of the wire is given by $J = J_0 \frac{r}{R}$, where J_0 is a constant. Find the magnetic field (a) at a point outside the wire and (b) at a point inside the wire. Write your answer in terms of the net current I through the wire.

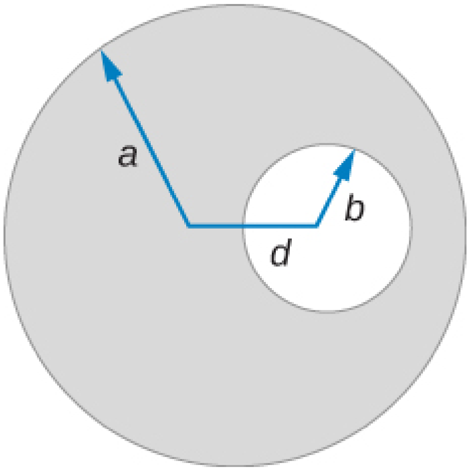
Solution:

a. $B = \frac{\mu_0 I}{2\pi r}$; b. $B = \frac{\mu_0 J_0 r^2}{3R}$

Exercise:

Problem:

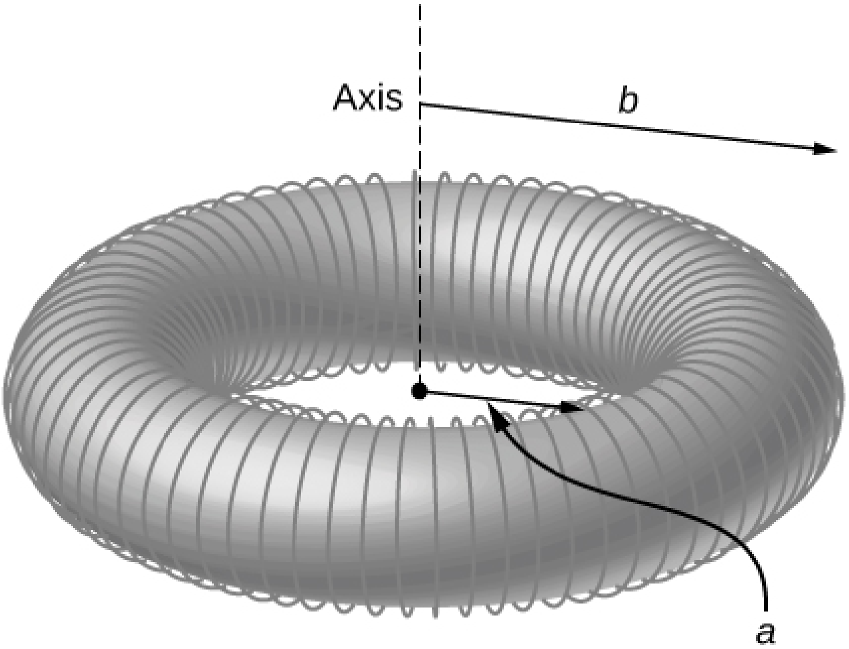
A very long, cylindrical wire of radius *a* has a circular hole of radius *b* in it at a distance *d* from the center. The wire carries a uniform current of magnitude *I* through it. The direction of the current in the figure is out of the paper. Find the magnetic field (a) at a point at the edge of the hole closest to the center of the thick wire, (b) at an arbitrary point inside the hole, and (c) at an arbitrary point outside the wire. (*Hint:* Think of the hole as a sum of two wires carrying current in the opposite directions.)



Exercise:

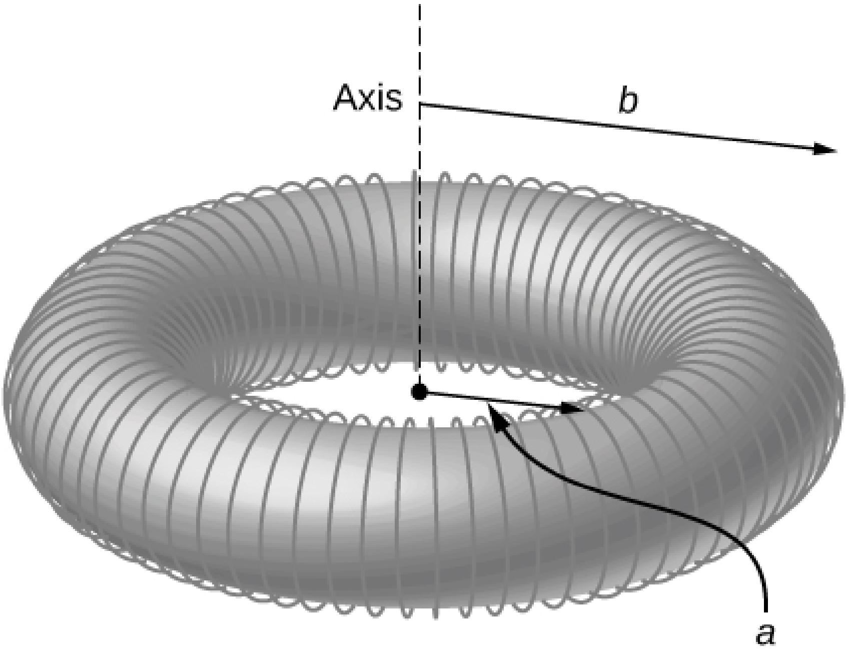
Problem:

Magnetic field inside a torus. Consider a torus of rectangular cross-section with inner radius *a* and outer radius *b*. *N* turns of an insulated thin wire are wound evenly on the torus tightly all around the torus and connected to a battery producing a steady current *I* in the wire. Assume that the current on the top and bottom surfaces in the figure is radial, and the current on the inner and outer radii surfaces is vertical. Find the magnetic field inside the torus as a function of radial distance *r* from the axis.



Solution:

$B(r) = \mu_0 NI/2\pi r$



Exercise:

Problem:

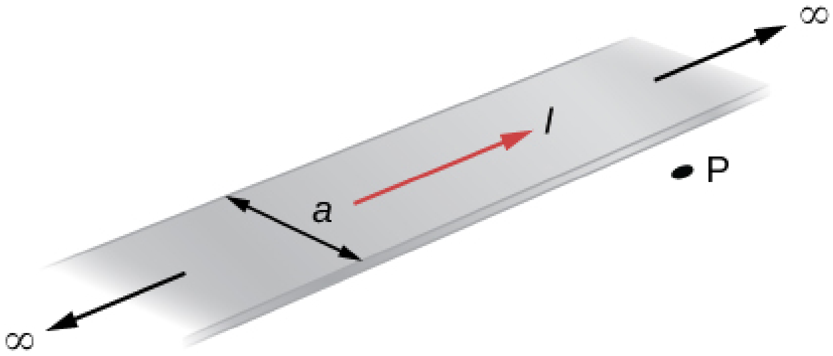
Two long coaxial copper tubes, each of length L, are connected to a battery of voltage V. The inner tube has inner radius a and outer radius b, and the outer tube has inner radius c and outer radius d. The tubes are then disconnected from the battery and rotated in the same direction at angular speed of ω radians per second about their common axis. Find the magnetic field (a) at a point inside the space enclosed by the inner tube $r < a$, and (b) at a point between the tubes $b < r < c$, and (c) at a point outside the tubes $r > d$. (*Hint:* Think of copper tubes as a capacitor and find the charge density based on the voltage applied, $Q = VC, C = \frac{2\pi\varepsilon_0 L}{\ln(c/b)}$.)

Challenge Problems

Exercise:

Problem:

The accompanying figure shows a flat, infinitely long sheet of width a that carries a current I uniformly distributed across it. Find the magnetic field at the point P, which is in the plane of the sheet and at a distance x from one edge. Test your result for the limit $a \to 0$.



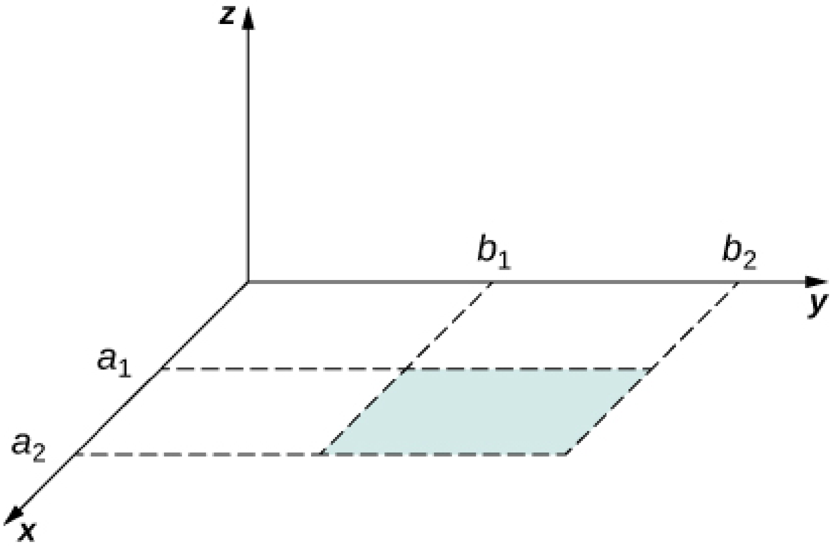
Solution:

$B = \frac{\mu_0 I}{2\pi a} \ln \frac{x+a}{x}$.

Exercise:

Problem:

A hypothetical current flowing in the *z*-direction creates the field $\vec{\mathbf{B}} = C\left[\left(x/y^2\right)\hat{\mathbf{i}} + (1/y)\hat{\mathbf{j}}\right]$ in the rectangular region of the *xy*-plane shown in the accompanying figure. Use Ampère’s law to find the current through the rectangle.



Exercise:

Problem:

A nonconducting hard rubber circular disk of radius *R* is painted with a uniform surface charge density σ. It is rotated about its axis with angular speed ω. (a) Find the magnetic field produced at a point on the axis a distance *h* meters from the center of the disk. (b) Find the numerical value of magnitude of the magnetic field when $\sigma = 1\text{C/m}^2$, $R = 20$ cm, $h = 2$ cm, and $\omega = 400\,\text{rad/sec}$, and compare it with the magnitude of magnetic field of Earth, which is about 1/2 Gauss.

Solution:

a. $B = \frac{\mu_0\sigma\omega}{2}\left[\frac{2h^2+R^2}{\sqrt{R^2+h^2}}-2h\right]$; b. $B = 4.09 \times 10^{-5}\text{T}$, 82% of Earth’s magnetic field

Glossary

diamagnetic materials

their magnetic dipoles align oppositely to an applied magnetic field; when the field is removed, the material is unmagnetized

ferromagnetic materials

contain groups of dipoles, called domains, that align with the applied magnetic field; when this field is removed, the material is still magnetized

hysteresis

property of ferromagnets that is seen when a material’s magnetic field is examined versus the applied magnetic field; a loop is created resulting from sweeping the applied field forward and reverse

magnetic susceptibility

ratio of the magnetic field in the material over the applied field at that time; positive susceptibilities are either paramagnetic or ferromagnetic (aligned with the field) and negative susceptibilities are diamagnetic (aligned oppositely with the field)

magnetic domains

groups of magnetic dipoles that are all aligned in the same direction and are coupled together quantum mechanically

paramagnetic materials

their magnetic dipoles align partially in the same direction as the applied magnetic field; when this field is removed, the material is unmagnetized

Introduction class="introduction"

The black strip found on the back of credit cards and driver’s licenses is a very thin layer of magnetic material with information stored on it. Reading and writing the information on the credit card is done with a swiping motion. The physical reason why this is necessary is called electromagnetic induction and is discussed in this chapter. (credit: modification of work by Jane Whitney)

www.cibc.com
Credit Card Services: 1-800-4
THIS CARD IS ISSUED BY AND IS
THE PROPERTY OF CIBC AND
MUST BE RETURNED TO CIBC
UPON REQUEST.
PLUS
Interac

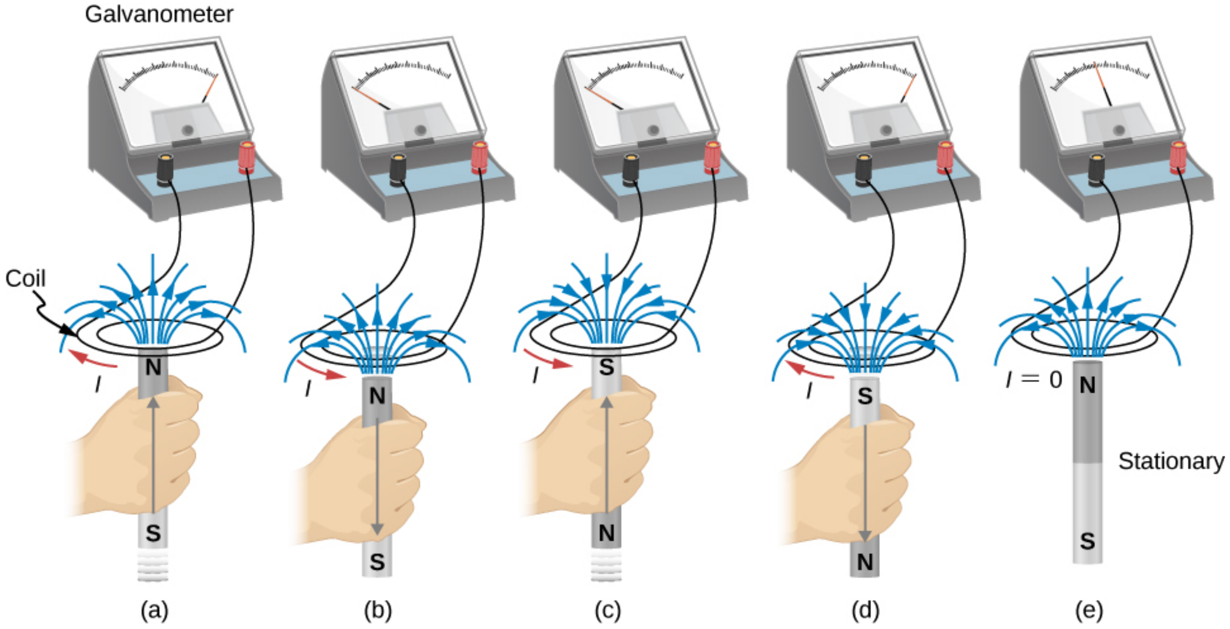
We have been considering electric fields created by fixed charge distributions and magnetic fields produced by constant currents, but electromagnetic phenomena are not restricted to these stationary situations. Most of the interesting applications of electromagnetism are, in fact, time-dependent. To investigate some of these applications, we now remove the time-independent assumption that we have been making and allow the fields to vary with time. In this and the next several chapters, you will see a wonderful symmetry in the behavior exhibited by time-varying electric and magnetic fields. Mathematically, this symmetry is expressed by an additional term in Ampère’s law and by another key equation of electromagnetism called Faraday’s law. We also discuss how moving a wire through a magnetic field produces an emf or voltage. Lastly, we describe applications of these principles, such as the card reader shown above.

Faraday’s Law

By the end of this section, you will be able to:

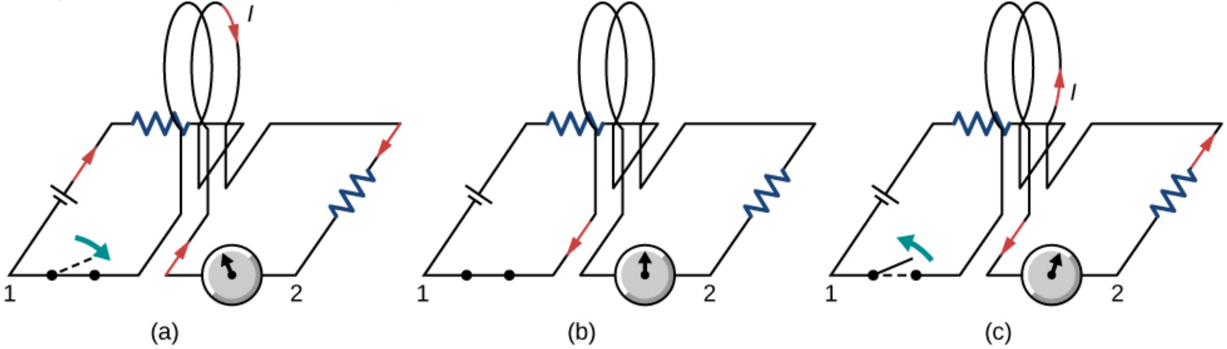
- Determine the magnetic flux through a surface, knowing the strength of the magnetic field, the surface area, and the angle between the normal to the surface and the magnetic field
- Use Faraday’s law to determine the magnitude of induced emf in a closed loop due to changing magnetic flux through the loop

The first productive experiments concerning the effects of time-varying magnetic fields were performed by Michael Faraday in 1831. One of his early experiments is represented in [link]. An emf is induced when the magnetic field in the coil is changed by pushing a bar magnet into or out of the coil. Emfs of opposite signs are produced by motion in opposite directions, and the directions of emfs are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet—it is the relative motion that is important. The faster the motion, the greater the emf, and there is no emf when the magnet is stationary relative to the coil.



Movement of a magnet relative to a coil produces emfs as shown (a–d). The same emfs are produced if the coil is moved relative to the magnet. This short-lived emf is only present during the motion. The greater the speed, the greater the magnitude of the emf, and the emf is zero when there is no motion, as shown in (e).

Faraday also discovered that a similar effect can be produced using two circuits—a changing current in one circuit induces a current in a second, nearby circuit. For example, when the switch is closed in circuit 1 of [link] (a), the ammeter needle of circuit 2 momentarily deflects, indicating that a short-lived current surge has been induced in that circuit. The ammeter needle quickly returns to its original position, where it remains. However, if the switch of circuit 1 is now suddenly opened, another short-lived current surge in the direction opposite from before is observed in circuit 2.



(a) Closing the switch of circuit 1 produces a short-lived current surge in circuit 2. (b) If the switch remains closed, no current is observed in circuit 2. (c) Opening the switch again produces a short-lived current in circuit 2 but in the opposite direction from before.

Faraday realized that in both experiments, a current flowed in the circuit containing the ammeter only when the magnetic field in the region occupied by that circuit was *changing*. As the magnet of the figure was moved, the strength of its magnetic field at the loop changed; and when the current in circuit 1 was turned on or off, the strength of its magnetic field at circuit 2 changed. Faraday was eventually able to interpret these and all other experiments involving magnetic fields that vary with time in terms of the following law:

Note:

Faraday’s Law

The emf ε induced is the negative change in the magnetic flux Φ_{m} per unit time. Any change in the magnetic field or change in orientation of the area of the coil with respect to the magnetic field induces a voltage (emf).

The **magnetic flux** is a measurement of the amount of magnetic field lines through a given surface area, as seen in [link]. This definition is similar to the electric flux studied earlier. This means that if we have

Note:

Equation:

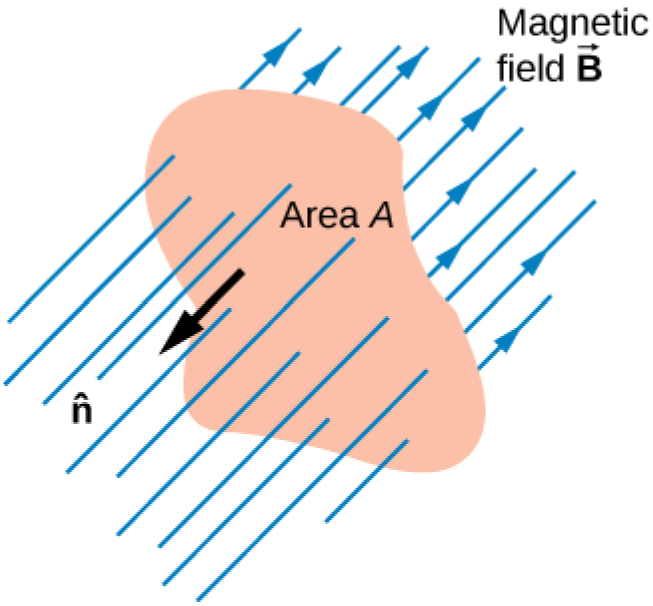
$$\Phi_{\mathrm{m}} = \int_S \vec{\mathbf{B}} \cdot \hat{\mathbf{n}} dA,$$

then the **induced emf** or the voltage generated by a conductor or coil moving in a magnetic field is

Equation:

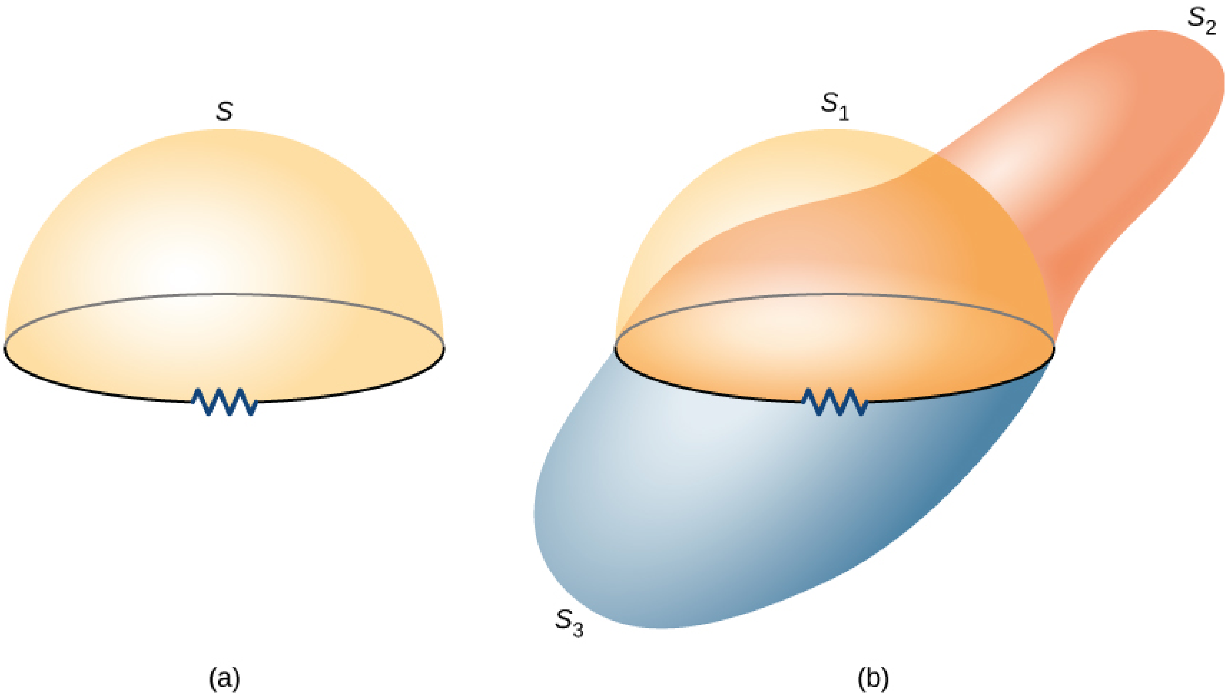
$$\varepsilon = -\frac{d}{dt} \int_S \vec{\mathbf{B}} \cdot \hat{\mathbf{n}} dA = -\frac{d\Phi_{\mathrm{m}}}{dt}.$$

The negative sign describes the direction in which the induced emf drives current around a circuit. However, that direction is most easily determined with a rule known as Lenz’s law, which we will discuss shortly.



The magnetic flux is the amount of magnetic field lines cutting through a surface area A defined by the unit area vector $\hat{\mathbf{n}}$. If the angle between the unit area $\hat{\mathbf{n}}$ and magnetic field vector $\vec{\mathbf{B}}$ are parallel or antiparallel, as shown in the diagram, the magnetic flux is the highest possible value given the values of area and magnetic field.

Part (a) of [link] depicts a circuit and an arbitrary surface S that it bounds. Notice that S is an *open surface*. It can be shown that *any* open surface bounded by the circuit in question can be used to evaluate Φ_{m}. For example, Φ_{m} is the same for the various surfaces $S_1, S_2, \ldots$ of part (b) of the figure.



(a) A circuit bounding an arbitrary open surface *S*. The planar area bounded by the circuit is not part of *S*. (b) Three arbitrary open surfaces bounded by the same circuit. The value of Φ_{m} is the same for all these surfaces.

The SI unit for magnetic flux is the weber (Wb),

Equation:

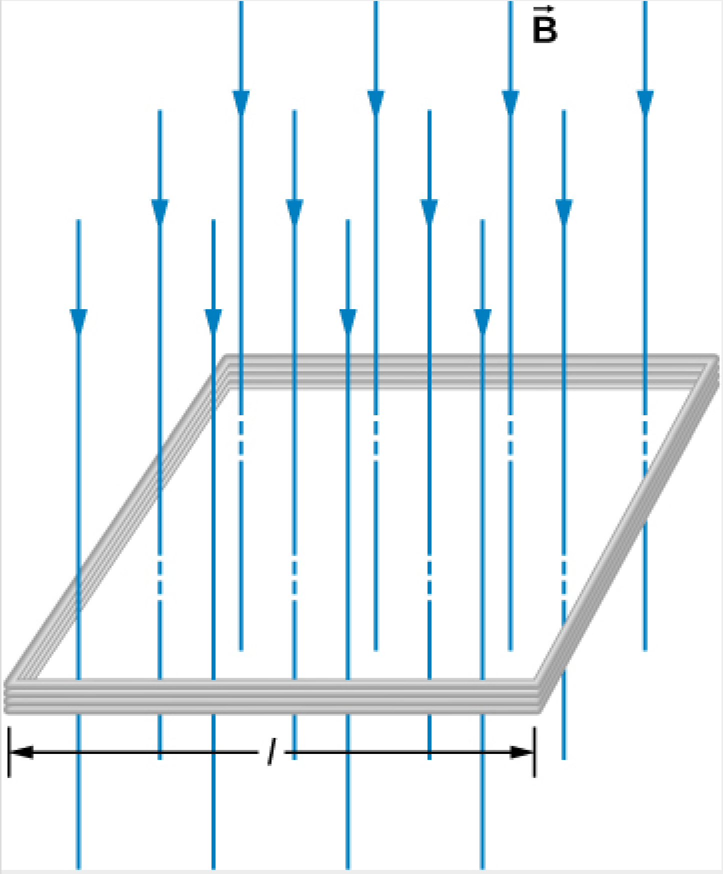
$$1\,\mathrm{Wb} = 1\,\mathrm{T} \cdot \mathrm{m}^2.$$

Occasionally, the magnetic field unit is expressed as webers per square meter $(\mathrm{Wb/m^2})$ instead of teslas, based on this definition. In many practical applications, the circuit of interest consists of a number *N* of tightly wound turns (see [link]). Each turn experiences the same magnetic flux. Therefore, the net magnetic flux through the circuits is *N* times the flux through one turn, and Faraday’s law is written as

Note:
Equation:

$$\varepsilon = -\frac{d}{dt}(N\Phi_\text{m}) = -N\frac{d\Phi_\text{m}}{dt}.$$

Example:
A Square Coil in a Changing Magnetic Field
The square coil of [link] has sides $l = 0.25\ \text{m}$ long and is tightly wound with $N = 200$ turns of wire. The resistance of the coil is $R = 5.0\ \Omega$. The coil is placed in a spatially uniform magnetic field that is directed perpendicular to the face of the coil and whose magnitude is decreasing at a rate $dB/dt = -0.040\ \text{T/s}$. (a) What is the magnitude of the emf induced in the coil? (b) What is the magnitude of the current circulating through the coil?



A square coil with *N* turns of wire with uniform magnetic field $\vec{\mathbf{B}}$ directed in the downward direction, perpendicular to the coil.

Strategy

The area vector, or $\widehat{\mathbf{n}}$ direction, is perpendicular to area covering the loop. We will choose this to be pointing downward so that $\vec{\mathbf{B}}$ is parallel to $\widehat{\mathbf{n}}$ and that the flux turns into multiplication of magnetic field times area. The area of the loop is not changing in time, so it can be factored out of the time derivative, leaving the magnetic field as the only quantity varying in time. Lastly, we can apply Ohm's law once we know the induced emf to find the current in the loop.

Solution

a. The flux through one turn is

Equation:

$$\Phi_{\mathrm{m}} = BA = Bl^2,$$

so we can calculate the magnitude of the emf from Faraday's law. The sign of the emf will be discussed in the next section, on Lenz's law:

Equation:

$$\begin{aligned} |\varepsilon| &= \left|-N\frac{d\Phi_{\mathrm{m}}}{dt}\right| = Nl^2\frac{dB}{dt} \\ &= (200)(0.25\ \mathrm{m})^2(0.040\ \mathrm{T/s}) = 0.50\ \mathrm{V}. \end{aligned}$$

b. The magnitude of the current induced in the coil is

Equation:

$$I = \frac{\varepsilon}{R} = \frac{0.50\ \mathrm{V}}{5.0\ \Omega} = 0.10\ \mathrm{A}.$$

Significance

If the area of the loop were changing in time, we would not be able to pull it out of the time derivative. Since the loop is a closed path, the result of this current would be a small amount of heating of the wires until the magnetic field stops changing. This may increase the area of the loop slightly as the wires are heated.

Note:
Exercise:

Problem:

Check Your Understanding A closely wound coil has a radius of 4.0 cm, 50 turns, and a total resistance of $40\ \Omega$. At what rate must a magnetic field perpendicular to the face of the coil change in order to produce Joule heating in the coil at a rate of 2.0 mW?

Solution:

1.1 T/s

Summary

- The magnetic flux through an enclosed area is defined as the amount of field lines cutting through a surface area *A* defined by the unit area vector.
- The units for magnetic flux are webers, where $1\ \text{Wb} = 1\ \text{T} \cdot \text{m}^2$.
- The induced emf in a closed loop due to a change in magnetic flux through the loop is known as Faraday’s law. If there is no change in magnetic flux, no induced emf is created.

Conceptual Questions

Exercise:

Problem:

A stationary coil is in a magnetic field that is changing with time. Does the emf induced in the coil depend on the actual values of the magnetic field?

Solution:

The emf depends on the rate of change of the magnetic field.

Exercise:

Problem:

In Faraday's experiments, what would be the advantage of using coils with many turns?

Exercise:

Problem:

A copper ring and a wooden ring of the same dimensions are placed in magnetic fields so that there is the same change in magnetic flux through them. Compare the induced electric fields and currents in the rings.

Solution:

Both have the same induced electric fields; however, the copper ring has a much higher induced emf because it conducts electricity better than the wooden ring.

Exercise:

Problem:

Discuss the factors determining the induced emf in a closed loop of wire.

Exercise:

Problem:

(a) Does the induced emf in a circuit depend on the resistance of the circuit? (b) Does the induced current depend on the resistance of the circuit?

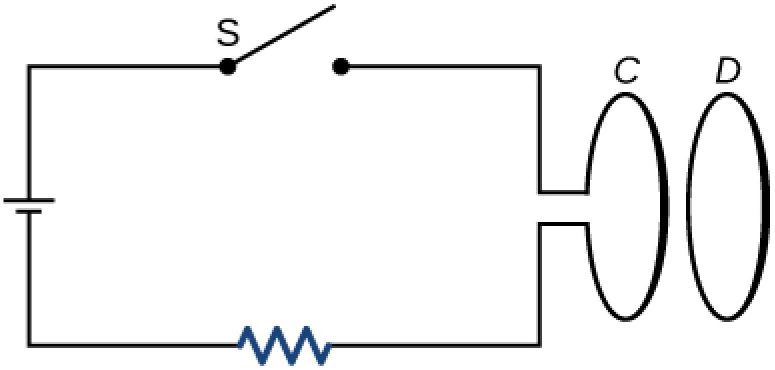
Solution:

a. no; b. yes

Exercise:

Problem:

How would changing the radius of loop *D* shown below affect its emf, assuming *C* and *D* are much closer together compared to their radii?



Exercise:

Problem:

Can there be an induced emf in a circuit at an instant when the magnetic flux through the circuit is zero?

Solution:

As long as the magnetic flux is changing from positive to negative or negative to positive, there could be an induced emf.

Exercise:

Problem:

Does the induced emf always act to decrease the magnetic flux through a circuit?

Exercise:

Problem:

How would you position a flat loop of wire in a changing magnetic field so that there is no induced emf in the loop?

Solution:

Position the loop so that the field lines run perpendicular to the area vector or parallel to the surface.

Exercise:

Problem:

The normal to the plane of a single-turn conducting loop is directed at an angle θ to a spatially uniform magnetic field $\vec{\mathbf{B}}$. It has a fixed area and orientation relative to the magnetic field. Show that the emf induced in the loop is given by $\varepsilon = (dB/dt)(A\cos\theta)$, where A is the area of the loop.

Problems

Exercise:

Problem:

A 50-turn coil has a diameter of 15 cm. The coil is placed in a spatially uniform magnetic field of magnitude 0.50 T so that the face of the coil and the magnetic field are perpendicular. Find the magnitude of the emf induced in the coil if the magnetic field is reduced to zero uniformly in (a) 0.10 s, (b) 1.0 s, and (c) 60 s.

Exercise:

Problem:

Repeat your calculations of the preceding problem’s time of 0.1 s with the plane of the coil making an angle of (a) 30°, (b) 60°, and (c) 90° with the magnetic field.

Solution:

a. 3.8 V; b. 2.2 V; c. 0 V

Exercise:

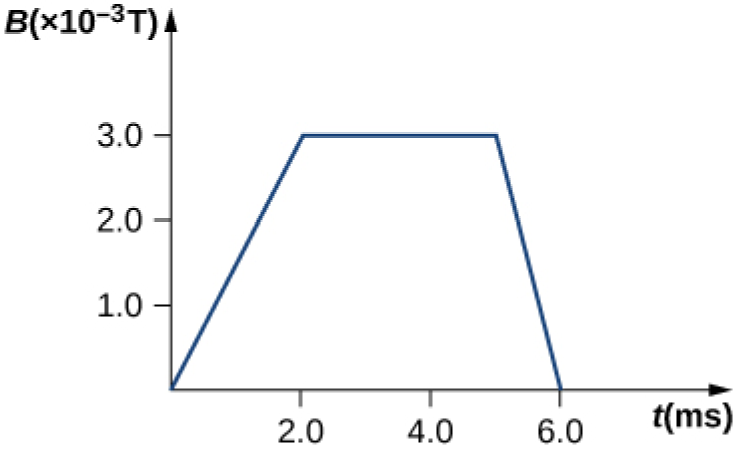
Problem:

A square loop whose sides are 6.0-cm long is made with copper wire of radius 1.0 mm. If a magnetic field perpendicular to the loop is changing at a rate of 5.0 mT/s, what is the current in the loop?

Exercise:

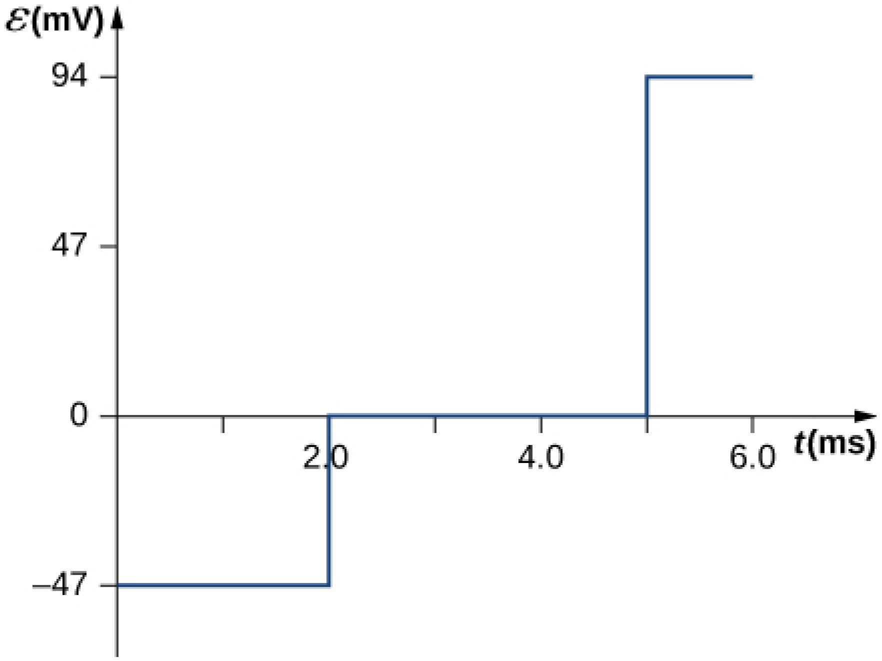
Problem:

The magnetic field through a circular loop of radius 10.0 cm varies with time as shown below. The field is perpendicular to the loop. Plot the magnitude of the induced emf in the loop as a function of time.



Solution:

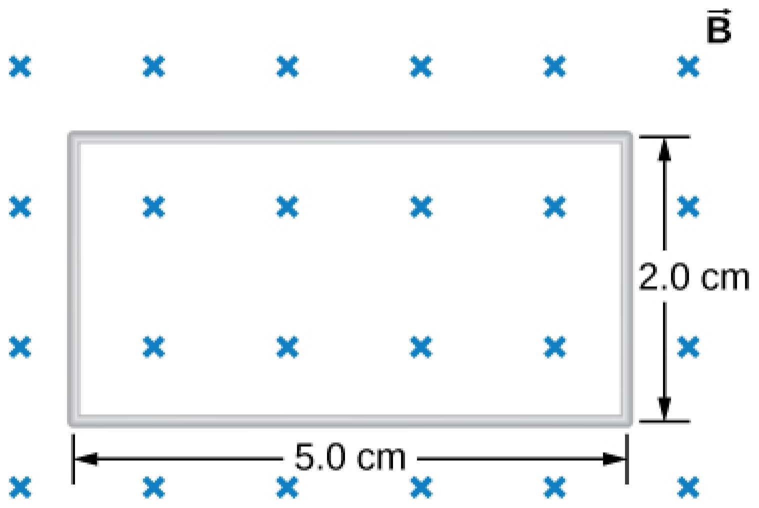
$$\begin{aligned}
B &= 1.5t, 0 \leq t < 2.0\,\text{ms}, B = 3.0\,\text{mT}, 2.0\,\text{ms} \leq t \leq 5.0\,\text{ms}, \\
B &= -3.0t + 18\,\text{mT}, 5.0\,\text{ms} < t \leq 6.0\,\text{ms}, \\
\varepsilon &= -\tfrac{d\Phi_\text{m}}{dt} = -\tfrac{d(BA)}{dt} = -A\tfrac{dB}{dt}, \\
\varepsilon &= -\pi(0.100\,\text{m})^2\,(1.5\,\text{T/s}) \\
&= -47\,\text{mV}\,(0 \leq t < 2.0\,\text{ms}), \\
\varepsilon &= \pi(0.100\,\text{m})^2\,(0) = 0\,(2.0\,\text{ms} \leq t \leq 5.0\,\text{ms}), \\
\varepsilon &= -\pi(0.100\,\text{m})^2\,(-3.0\,\text{T/s}) = 94\,\text{mV}\;(5.0\,\text{ms} < t < 6.0\,\text{ms}).
\end{aligned}$$



Exercise:

Problem:

The accompanying figure shows a single-turn rectangular coil that has a resistance of $2.0\,\Omega$. The magnetic field at all points inside the coil varies according to $B = B_0 e^{-\alpha t}$, where $B_0 = 0.25\,\text{T}$ and $\alpha = 200\,\text{Hz}$. What is the current induced in the coil at (a) $t = 0.001\,\text{s}$, (b) 0.002 s, (c) 2.0 s?



Exercise:

Problem:

How would the answers to the preceding problem change if the coil consisted of 20 closely spaced turns?

Solution:

Each answer is 20 times the previously given answers.

Exercise:

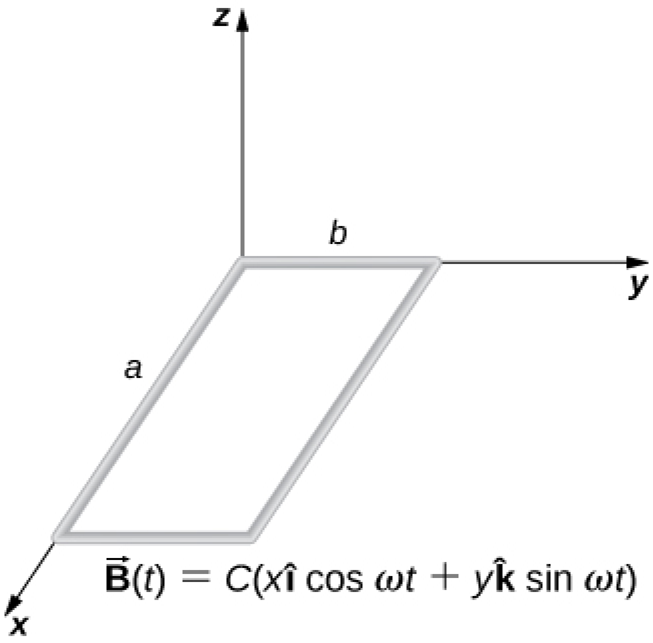
Problem:

A long solenoid with $n = 10$ turns per centimeter has a cross-sectional area of $5.0\ \text{cm}^2$ and carries a current of 0.25 A. A coil with five turns encircles the solenoid. When the current through the solenoid is turned off, it decreases to zero in 0.050 s. What is the average emf induced in the coil?

Exercise:

Problem:

A rectangular wire loop with length *a* and width *b* lies in the *xy*-plane, as shown below. Within the loop there is a time-dependent magnetic field given by $\vec{\mathbf{B}}(t) = C\left((x\cos\omega t)\hat{\mathbf{i}} + (y\sin\omega t)\widehat{\mathbf{k}}\right)$, with $\vec{\mathbf{B}}(t)$ in tesla. Determine the emf induced in the loop as a function of time.


z
b
y
a
x
$\vec{\mathbf{B}}(t) = C(x\hat{\mathbf{i}}\cos\omega t + y\hat{\mathbf{k}}\sin\omega t)$

Solution:

$$\begin{aligned} \hat{\mathbf{n}} &= \widehat{k},\ d\Phi_{\mathrm{m}} = \mathrm{C}y\sin(\omega t)dxdy, \\ \Phi_{\mathrm{m}} &= \frac{Cab^2\sin(\omega t)}{2}, \\ \varepsilon &= -\frac{Cab^2\omega\cos(\omega t)}{2}. \end{aligned}$$

Exercise:

Problem:

The magnetic field perpendicular to a single wire loop of diameter 10.0 cm decreases from 0.50 T to zero. The wire is made of copper and has a diameter of 2.0 mm and length 1.0 cm. How much charge moves through the wire while the field is changing?

Glossary

Faraday’s law
    induced emf is created in a closed loop due to a change in magnetic flux through the loop

induced emf
    short-lived voltage generated by a conductor or coil moving in a magnetic field

magnetic flux
    measurement of the amount of magnetic field lines through a given area

Lenz's Law

By the end of this section, you will be able to:

- Use Lenz’s law to determine the direction of induced emf whenever a magnetic flux changes
- Use Faraday’s law with Lenz’s law to determine the induced emf in a coil and in a solenoid

The direction in which the induced emf drives current around a wire loop can be found through the negative sign. However, it is usually easier to determine this direction with **Lenz’s law**, named in honor of its discoverer, Heinrich Lenz (1804–1865). (Faraday also discovered this law, independently of Lenz.) We state Lenz’s law as follows:

Note:

Lenz’s Law

The direction of the induced emf drives current around a wire loop to always *oppose* the change in magnetic flux that causes the emf.

Lenz’s law can also be considered in terms of conservation of energy. If pushing a magnet into a coil causes current, the energy in that current must have come from somewhere. If the induced current causes a magnetic field opposing the increase in field of the magnet we pushed in, then the situation is clear. We pushed a magnet against a field and did work on the system, and that showed up as current. If it were not the case that the induced field opposes the change in the flux, the magnet would be pulled in produce a current without anything having done work. Electric potential energy would have been created, violating the conservation of energy.

To determine an induced emf ε, you first calculate the magnetic flux Φ_{m} and then obtain $d\Phi_{\mathrm{m}}/dt$. The magnitude of ε is given by $\varepsilon = |d\Phi_{\mathrm{m}}/dt|$. Finally, you can apply Lenz’s law to determine the sense of ε. This will be developed through examples that illustrate the following problem-solving strategy.

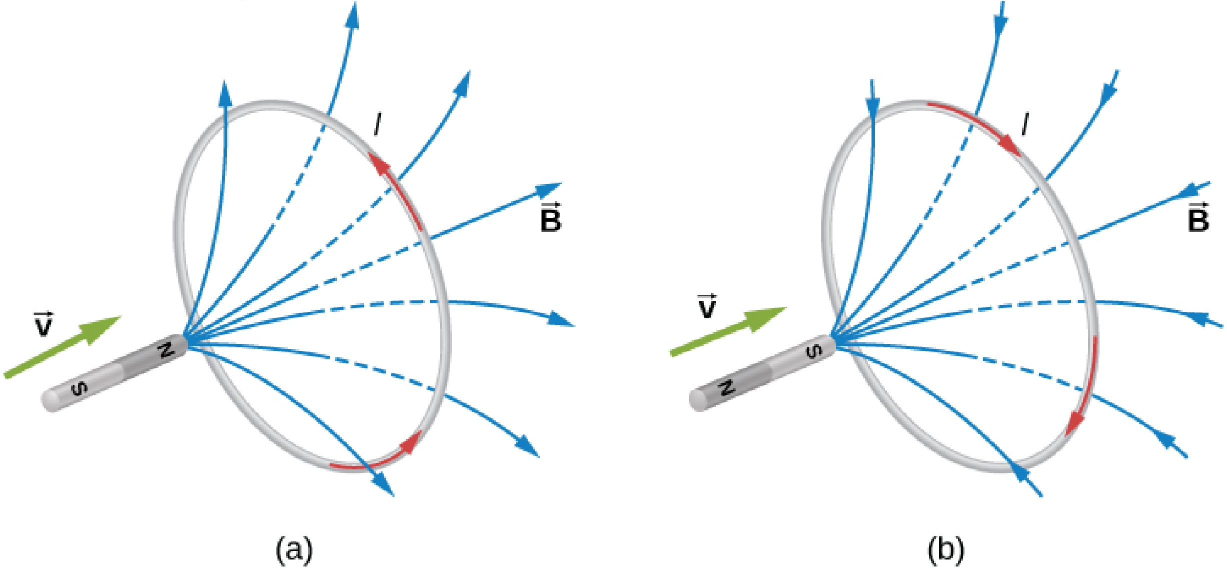
Note:

Problem-Solving Strategy: Lenz’s Law

To use Lenz’s law to determine the directions of induced magnetic fields, currents, and emfs:

1. Make a sketch of the situation for use in visualizing and recording directions.
2. Determine the direction of the applied magnetic field $\vec{\mathbf{B}}$.
3. Determine whether its magnetic flux is increasing or decreasing.
4. Now determine the direction of the induced magnetic field $\vec{\mathbf{B}}$. The induced magnetic field tries to reinforce a magnetic flux that is decreasing or opposes a magnetic flux that is increasing. Therefore, the induced magnetic field adds or subtracts to the applied magnetic field, depending on the change in magnetic flux.
5. Use right-hand rule 2 (RHR-2; see Magnetic Forces and Fields) to determine the direction of the induced current I that is responsible for the induced magnetic field $\vec{\mathbf{B}}$.
6. The direction (or polarity) of the induced emf can now drive a conventional current in this direction.

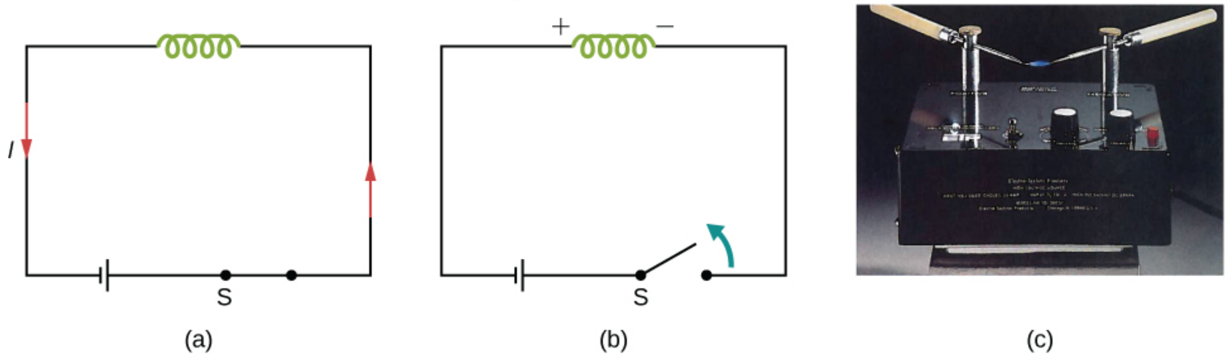
Let’s apply Lenz’s law to the system of [link](a). We designate the “front” of the closed conducting loop as the region containing the approaching bar magnet, and the “back” of the loop as the other region. As the north pole of the magnet moves toward the loop, the flux through the loop due to the field of the magnet increases because the strength of field lines directed from the front to the back of the loop is increasing. A current is therefore induced in the loop. By Lenz’s law, the direction of the induced current must be such that its own magnetic field is directed in a way to *oppose* the changing flux caused by the field of the approaching magnet. Hence, the induced current circulates so that its magnetic field lines through the loop are directed from the back to the front of the loop. By RHR-2, place your thumb pointing against the magnetic field lines, which is toward the bar magnet. Your fingers wrap in a counterclockwise direction as viewed from the bar magnet. Alternatively, we can determine the direction of the induced current by treating the current loop as an electromagnet that *opposes* the approach of the north pole of the bar magnet. This occurs when the induced current flows as shown, for then the face of the loop nearer the approaching magnet is also a north pole.



The change in magnetic flux caused by the approaching magnet induces a current in the loop. (a) An approaching north pole induces a counterclockwise current with respect to the bar magnet. (b) An approaching south pole induces a clockwise current with respect to the bar magnet.

Part (b) of the figure shows the south pole of a magnet moving toward a conducting loop. In this case, the flux through the loop due to the field of the magnet increases because the number of field lines directed from the back to the front of the loop is increasing. To oppose this change, a current is induced in the loop whose field lines through the loop are directed from the front to the back. Equivalently, we can say that the current flows in a direction so that the face of the loop nearer the approaching magnet is a south pole, which then repels the approaching south pole of the magnet. By RHR-2, your thumb points away from the bar magnet. Your fingers wrap in a clockwise fashion, which is the direction of the induced current.

Another example illustrating the use of Lenz’s law is shown in [link]. When the switch is opened, the decrease in current through the solenoid causes a decrease in magnetic flux through its coils, which induces an emf in the solenoid. This emf must oppose the change (the termination of the current) causing it. Consequently, the induced emf has the polarity shown and drives in the direction of the original current. This may generate an arc across the terminals of the switch as it is opened.



(a) A solenoid connected to a source of emf. (b) Opening switch S terminates the current, which in turn induces an emf in the solenoid. (c) A potential difference between the ends of the sharply pointed rods is produced by inducing an emf in a coil. This potential difference is large enough to produce an arc between the sharp points.

Note:

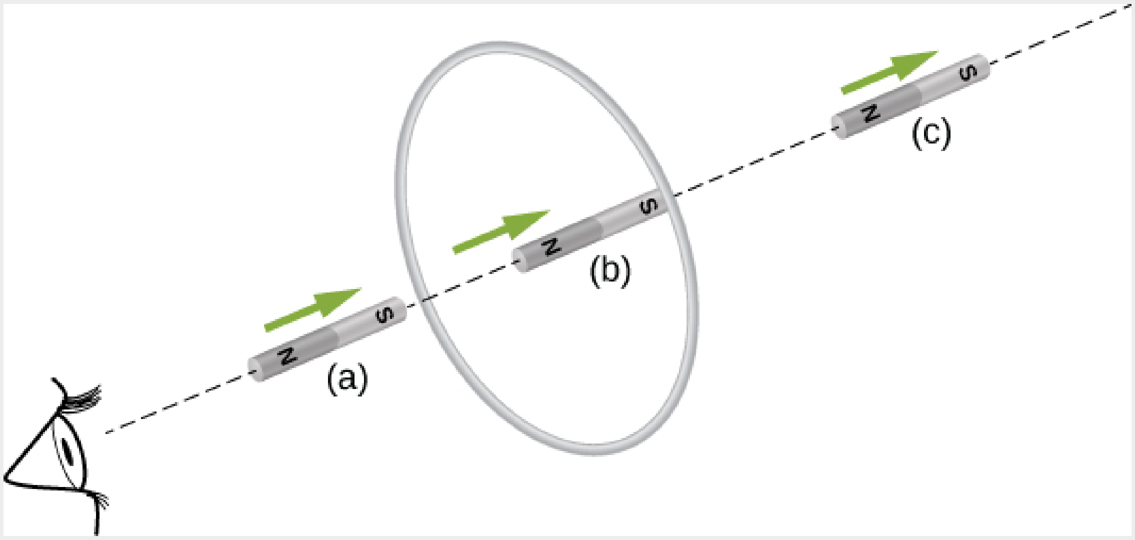
Exercise:

Problem:

Check Your Understanding Find the direction of the induced current in the wire loop shown below as the magnet enters, passes through, and leaves the loop.

Solution:

To the observer shown, the current flows clockwise as the magnet approaches, decreases to zero when the magnet is centered in the plane of the coil, and then flows counterclockwise as the magnet leaves the coil.



Note:
Exercise:

Problem:

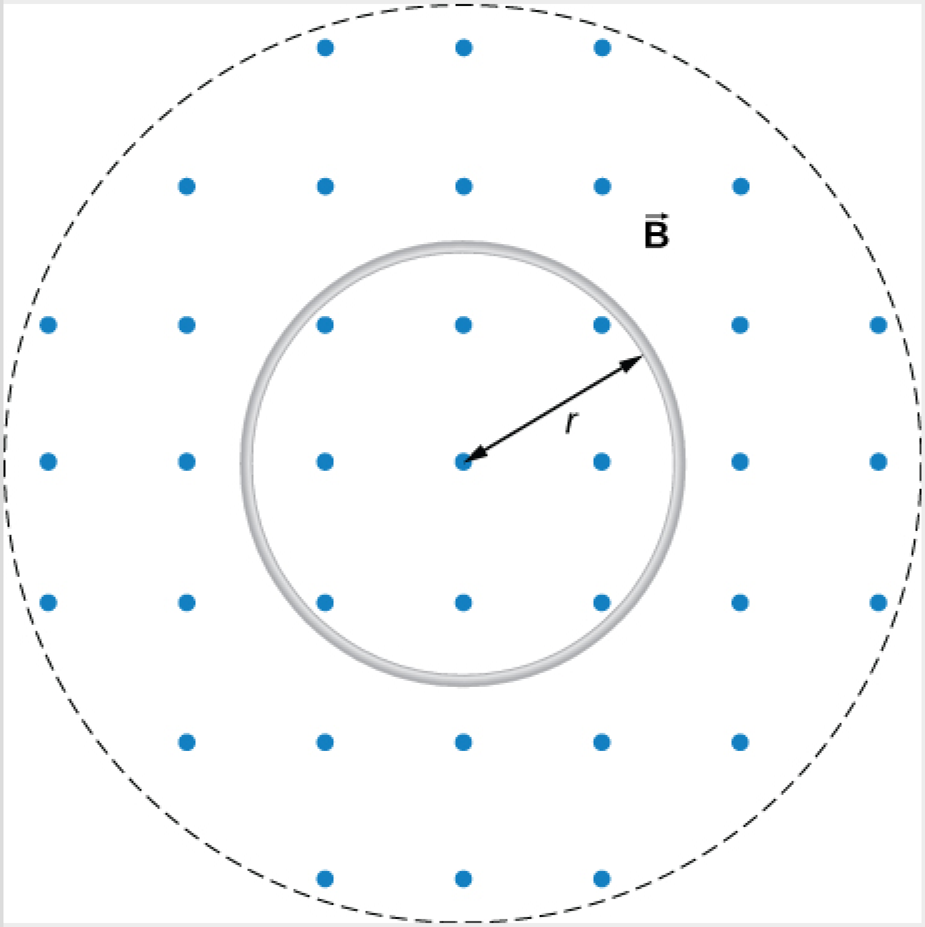
Check Your Understanding Verify the directions of the induced currents in [link].

Example:

A Circular Coil in a Changing Magnetic Field

A magnetic field $\vec{\mathbf{B}}$ is directed outward perpendicular to the plane of a circular coil of radius $r = 0.50\ \text{m}$ ([link]). The field is cylindrically symmetrical with respect to the center of the coil, and its magnitude decays exponentially according to $B = (1.5T)e^{-(5.0\text{s}^{-1})t}$, where B is in teslas and

t is in seconds. (a) Calculate the emf induced in the coil at the times $t_1 = 0$, $t_2 = 5.0 \times 10^{-2}$s, and $t_3 = 1.0\,\text{s}$. (b) Determine the current in the coil at these three times if its resistance is $10\,\Omega$.



A circular coil in a decreasing magnetic field.

Strategy

Since the magnetic field is perpendicular to the plane of the coil and constant over each spot in the coil, the dot product of the magnetic field $\vec{\mathbf{B}}$ and normal to the area unit vector $\hat{\mathbf{n}}$ turns into a multiplication. The magnetic field can be pulled out of the integration, leaving the flux as the product of the magnetic field times area. We need to take the time derivative of the exponential function to calculate the emf using Faraday's law. Then we use Ohm's law to calculate the current.

Solution

a. Since $\vec{\mathbf{B}}$ is perpendicular to the plane of the coil, the magnetic flux is given by
Equation:

$$\begin{aligned}\Phi_{\mathrm{m}} &= B\pi r^2 = (1.5e^{-5.0t}\ \mathrm{T})\pi(0.50\ \mathrm{m})^2\\ &= 1.2e^{-(5.0\mathrm{s}^{-1})t}\ \mathrm{Wb}.\end{aligned}$$

From Faraday’s law, the magnitude of the induced emf is
Equation:

$$\varepsilon = \left|\frac{d\Phi_m}{dt}\right| = \left|\frac{d}{dt}(1.2e^{-(5.0\mathrm{s}^{-1})t}\ \mathrm{Wb})\right| = 6.0\, e^{-(5.0\mathrm{s}^{-1})t}\ \mathrm{V}.$$

Since $\vec{\mathbf{B}}$ is directed out of the page and is decreasing, the induced current must flow counterclockwise when viewed from above so that the magnetic field it produces through the coil also points out of the page. For all three times, the sense of ε is counterclockwise; its magnitudes are
Equation:

$$\varepsilon\,(t_1) = 6.0\ \mathrm{V};\ \ \varepsilon\,(t_2) = 4.7\ \mathrm{V};\ \ \varepsilon\,(t_3) = 0.040\ \mathrm{V}.$$

b. From Ohm’s law, the respective currents are
Equation:

$$\begin{aligned}I(t_1) &= \frac{\varepsilon(t_1)}{R} = \frac{6.0\ \mathrm{V}}{10\ \Omega} = 0.60\ \mathrm{A};\\ I(t_2) &= \frac{4.7\ \mathrm{V}}{10\ \Omega} = 0.47\ \mathrm{A};\end{aligned}$$

and
Equation:

$$I(t_3) = \frac{0.040\ \mathrm{V}}{10\ \Omega} = 4.0\ \times\ 10^{-3}\ \mathrm{A}.$$

Significance

An emf voltage is created by a changing magnetic flux over time. If we know how the magnetic field varies with time over a constant area, we can take its time derivative to calculate the induced emf.

Example:

Changing Magnetic Field Inside a Solenoid

The current through the windings of a solenoid with $n = 2000$ turns per meter is changing at a rate $dI/dt = 3.0\ \text{A/s}$. (See Sources of Magnetic Fields for a discussion of solenoids.) The solenoid is 50-cm long and has a cross-sectional diameter of 3.0 cm. A small coil consisting of $N = 20$ closely wound turns wrapped in a circle of diameter 1.0 cm is placed in the middle of the solenoid such that the plane of the coil is perpendicular to the central axis of the solenoid. Assuming that the infinite-solenoid approximation is valid at the location of the small coil, determine the magnitude of the emf induced in the coil.

Strategy

The magnetic field in the middle of the solenoid is a uniform value of $\mu_0 nI$. This field is producing a maximum magnetic flux through the coil as it is directed along the length of the solenoid. Therefore, the magnetic flux through the coil is the product of the solenoid's magnetic field times the area of the coil. Faraday's law involves a time derivative of the magnetic flux. The only quantity varying in time is the current, the rest can be pulled out of the time derivative. Lastly, we include the number of turns in the coil to determine the induced emf in the coil.

Solution

Since the field of the solenoid is given by $B = \mu_0 nI$, the flux through each turn of the small coil is

Equation:

$$\Phi_\text{m} = \mu_0 nI \quad \frac{\pi d^2}{4} \quad ,$$

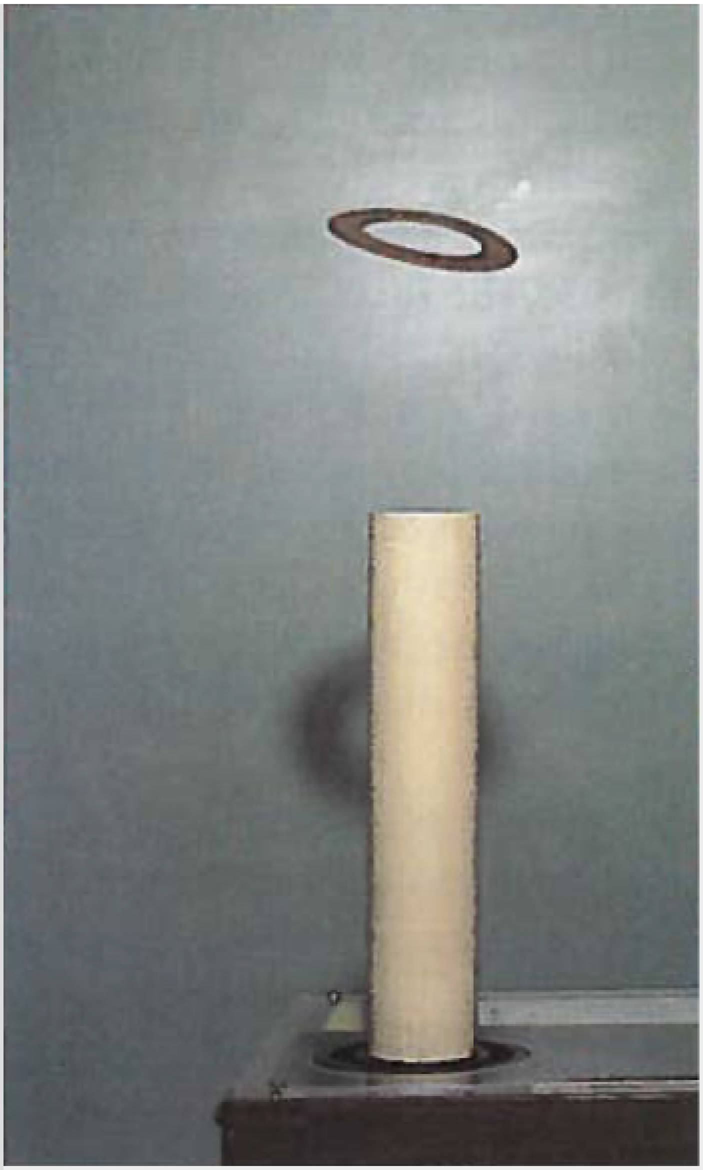
where *d* is the diameter of the coil. Now from Faraday’s law, the magnitude of the emf induced in the coil is

Equation:

$$\begin{aligned} \varepsilon &= \left| N \frac{d\Phi_{\mathrm{m}}}{dt} \right| = \left| N \mu_0 n \frac{\pi d^2}{4} \ \frac{dI}{dt} \right| \\ &= 20 \left(4\pi \ \times \ 10^{-7}\,\mathrm{T \cdot m/s}\right) \left(2000\,\mathrm{m}^{-1}\right) \frac{\pi (0.010\,\mathrm{m})^2}{4} (3.0\,\mathrm{A/s}) \\ &= 1.2 \ \times \ 10^{-5}\,\mathrm{V}. \end{aligned}$$

Significance

When the current is turned on in a vertical solenoid, as shown in [link], the ring has an induced emf from the solenoid’s changing magnetic flux that opposes the change. The result is that the ring is fired vertically into the air.



The jumping ring. When a current is turned on in the vertical solenoid, a current is induced in the metal ring. The stray field produced by the solenoid causes the ring to jump off the solenoid.

Note:
Visit this website for a demonstration of the jumping ring from MIT.

Summary

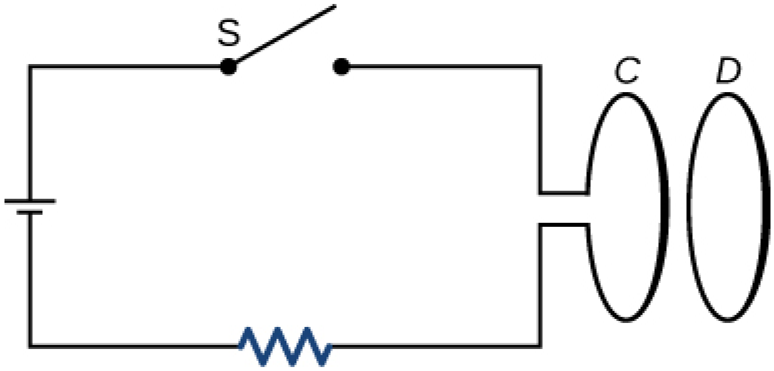
- We can use Lenz’s law to determine the directions of induced magnetic fields, currents, and emfs.
- The direction of an induced emf always opposes the change in magnetic flux that causes the emf, a result known as Lenz’s law.

Conceptual Questions

Exercise:

Problem:

The circular conducting loops shown in the accompanying figure are parallel, perpendicular to the plane of the page, and coaxial. (a) When the switch S is closed, what is the direction of the current induced in *D*? (b) When the switch is opened, what is the direction of the current induced in loop *D*?



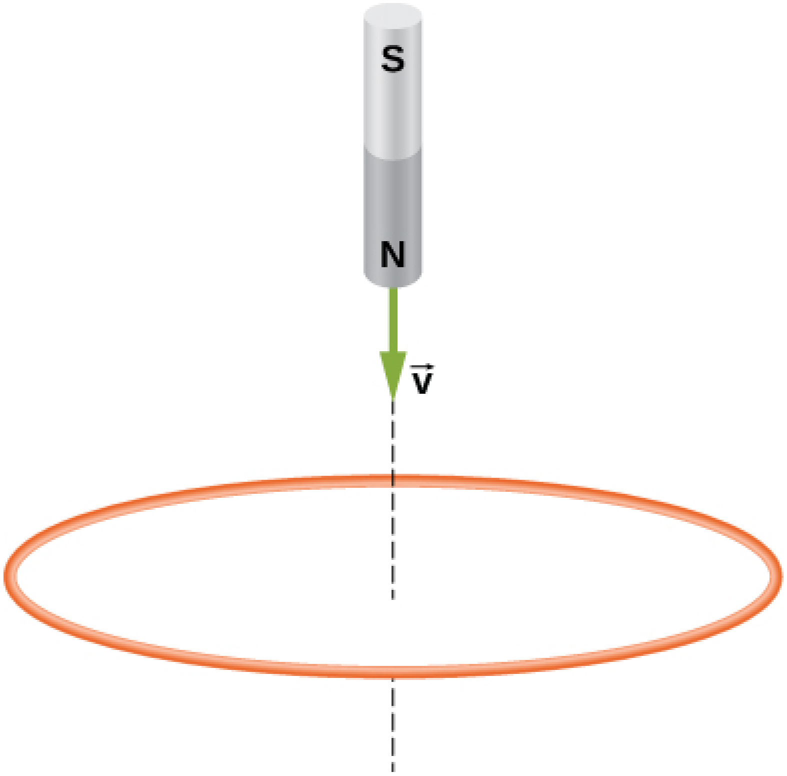
Solution:

a. CW as viewed from the circuit; b. CCW as viewed from the circuit

Exercise:

Problem:

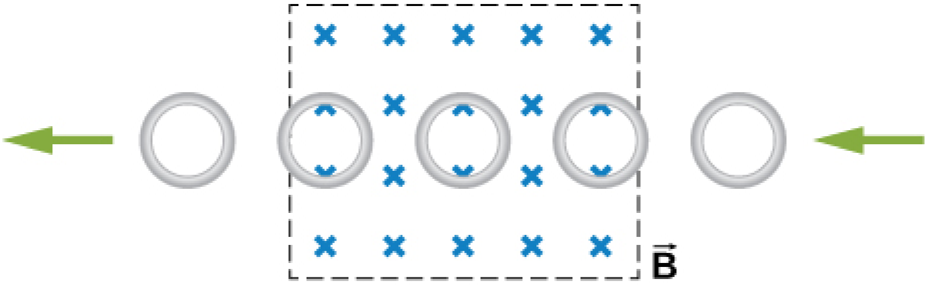
The north pole of a magnet is moved toward a copper loop, as shown below. If you are looking at the loop from above the magnet, will you say the induced current is circulating clockwise or counterclockwise?



Exercise:

Problem:

The accompanying figure shows a conducting ring at various positions as it moves through a magnetic field. What is the sense of the induced emf for each of those positions?



Solution:

As the loop enters, the induced emf creates a CCW current while as the loop leaves the induced emf creates a CW current. While the loop is fully inside the magnetic field, there is no flux change and therefore no induced current.

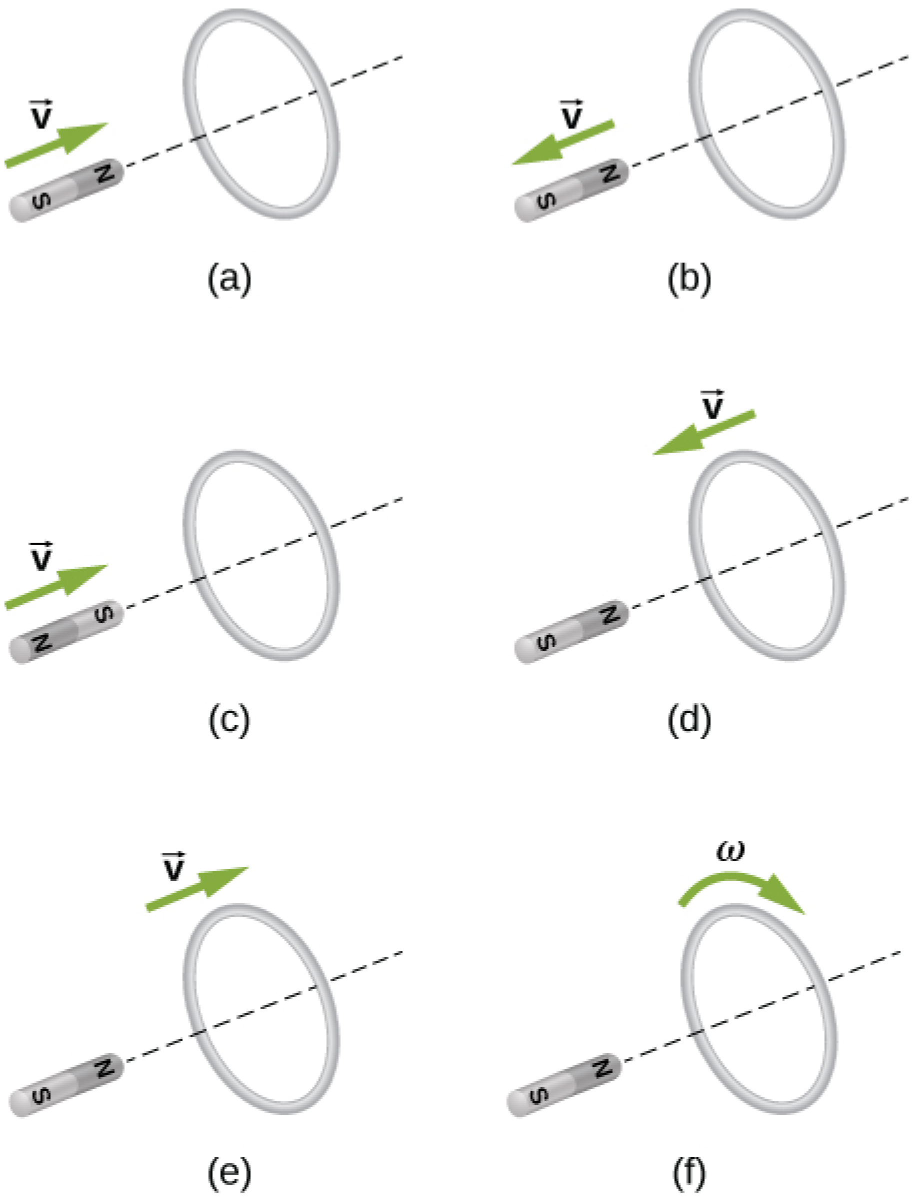
Exercise:

Problem: Show that ε and $d\Phi_{\mathrm{m}}/dt$ have the same units.

Exercise:

Problem:

State the direction of the induced current for each case shown below, observing from the side of the magnet.



(a) (b) (c) (d) (e) (f)

Solution:

a. CCW viewed from the magnet; b. CW viewed from the magnet; c. CW viewed from the magnet; d. CCW viewed from the magnet; e. CW viewed from the magnet; f. no current

Problems

Exercise:

Problem:

A single-turn circular loop of wire of radius 50 mm lies in a plane perpendicular to a spatially uniform magnetic field. During a 0.10-s time interval, the magnitude of the field increases uniformly from 200 to 300 mT. (a) Determine the emf induced in the loop. (b) If the magnetic field is directed out of the page, what is the direction of the current induced in the loop?

Solution:

a. 7.8×10^{-3}V; b. CCW from the same view as the magnetic field

Exercise:

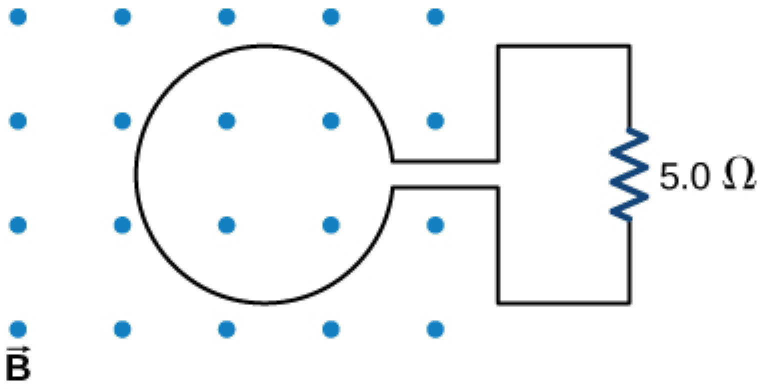
Problem:

When a magnetic field is first turned on, the flux through a 20-turn loop varies with time according to $\Phi_{\text{m}} = 5.0t^2 - 2.0t$, where Φ_{m} is in milliwebers, *t* is in seconds, and the loop is in the plane of the page with the unit normal pointing outward. (a) What is the emf induced in the loop as a function of time? What is the direction of the induced current at (b) *t* = 0, (c) 0.10, (d) 1.0, and (e) 2.0 s?

Exercise:

Problem:

The magnetic flux through the loop shown in the accompanying figure varies with time according to $\Phi_{\text{m}} = 2.00e^{-3t}\sin(120\pi t)$, where Φ_{m} is in milliwebers. What are the direction and magnitude of the current through the 5.00-Ω resistor at (a) $t = 0$; (b) $t = 2.17 \times 10^{-2}$ s, and (c) $t = 3.00$ s?



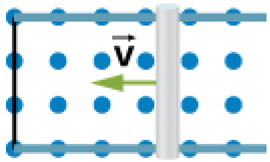
Solution:

a. 150 A downward through the resistor; b. 46 A upward through the resistor; c. 0.019 A downward through the resistor

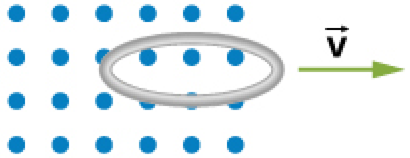
Exercise:

Problem:

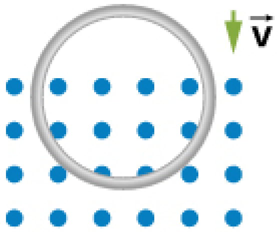
Use Lenz’s law to determine the direction of induced current in each case.



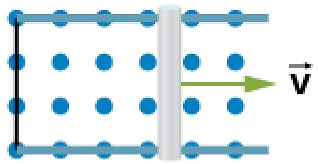
(a)



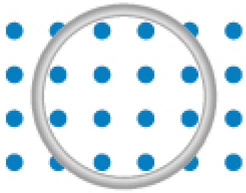
(b)



(c)

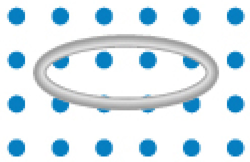


(d)



B increasing

(e)



B decreasing

(f)

Glossary

Lenz’s law

direction of an induced emf opposes the change in magnetic flux that produced it; this is the negative sign in Faraday’s law

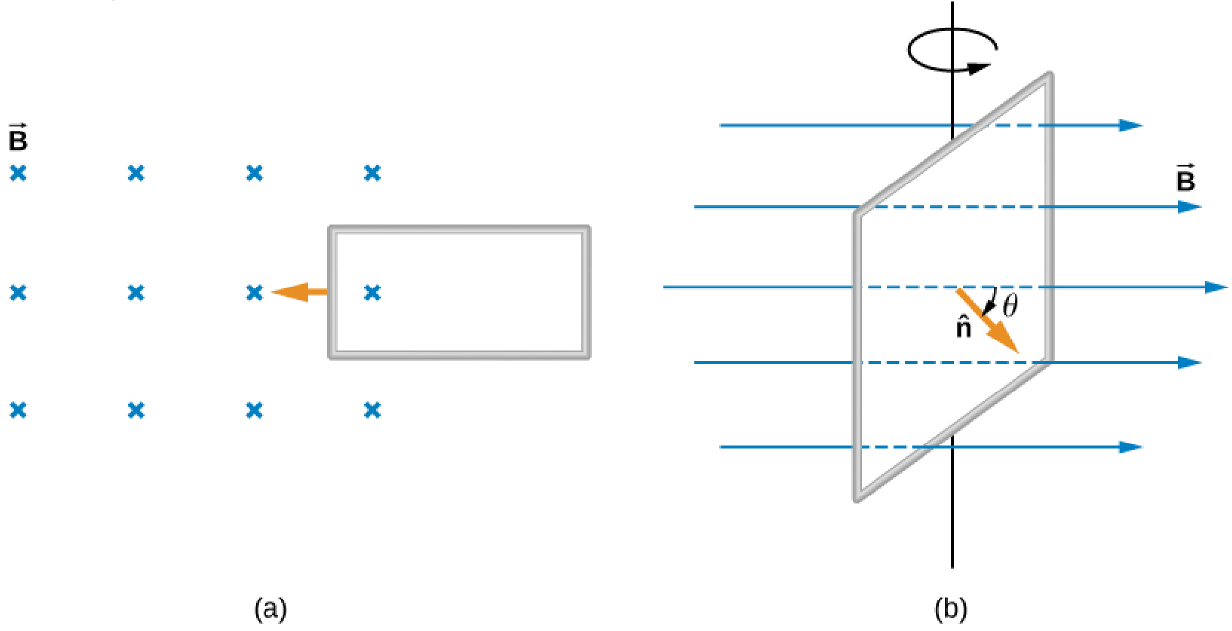
Motional Emf

By the end of this section, you will be able to:

- Determine the magnitude of an induced emf in a wire moving at a constant speed through a magnetic field
- Discuss examples that use motional emf, such as a rail gun and a tethered satellite

Magnetic flux depends on three factors: the strength of the magnetic field, the area through which the field lines pass, and the orientation of the field with the surface area. If any of these quantities varies, a corresponding variation in magnetic flux occurs. So far, we've only considered flux changes due to a changing field. Now we look at another possibility: a changing area through which the field lines pass including a change in the orientation of the area.

Two examples of this type of flux change are represented in [link]. In part (a), the flux through the rectangular loop increases as it moves into the magnetic field, and in part (b), the flux through the rotating coil varies with the angle θ.



(a) Magnetic flux changes as a loop moves into a magnetic field; (b) magnetic flux changes as a loop rotates in a magnetic field.

It’s interesting to note that what we perceive as the cause of a particular flux change actually depends on the frame of reference we choose. For example, if you are at rest relative to the moving coils of [link], you would see the flux vary because of a changing magnetic field—in part (a), the field moves from left to right in your reference frame, and in part (b), the field is rotating. It is often possible to describe a flux change through a coil that is moving in one particular reference frame in terms of a changing magnetic field in a second frame, where the coil is stationary. However, reference-frame questions related to magnetic flux are beyond the level of this textbook. We’ll avoid such complexities by always working in a frame at rest relative to the laboratory and explain flux variations as due to either a changing field or a changing area.

Now let’s look at a conducting rod pulled in a circuit, changing magnetic flux. The area enclosed by the circuit ‘MNOP’ of [link] is *lx* and is perpendicular to the magnetic field, so we can simplify the integration of [link] into a multiplication of magnetic field and area. The magnetic flux through the open surface is therefore

Equation:

$$\Phi_{\text{m}} = Blx.$$

Since *B* and *l* are constant and the velocity of the rod is $v = dx/dt$, we can now restate Faraday’s law, [link], for the magnitude of the emf in terms of the moving conducting rod as

Note:

Equation:

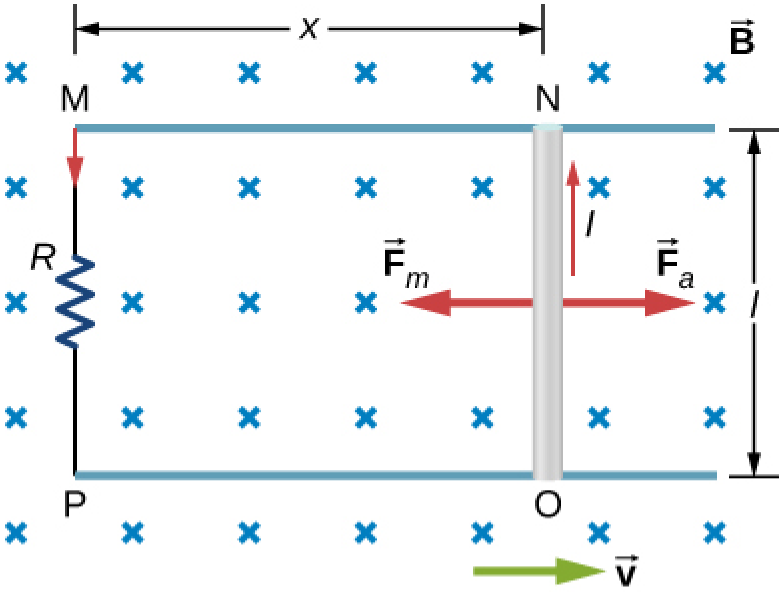
$$\varepsilon = \frac{d\Phi_{\text{m}}}{dt} = Bl\frac{dx}{dt} = Blv.$$

The current induced in the circuit is the emf divided by the resistance or
Equation:

$$I = \frac{Blv}{R}.$$

Furthermore, the direction of the induced emf satisfies Lenz's law, as you can verify by inspection of the figure.

This calculation of motionally induced emf is not restricted to a rod moving on conducting rails. With $\vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$ as the starting point, it can be shown that $\varepsilon = -d\Phi_{\mathrm{m}}/dt$ holds for any change in flux caused by the motion of a conductor. We saw in Faraday's Law that the emf induced by a time-varying magnetic field obeys this same relationship, which is Faraday's law. Thus Faraday's law *holds for all flux changes*, whether they are produced by a changing magnetic field, by motion, or by a combination of the two.



A conducting rod is pushed to the right at constant velocity. The resulting change in the magnetic flux induces a current in the circuit.

From an energy perspective, $\vec{\mathbf{F}}_a$ produces power $F_a v$, and the resistor dissipates power $I^2 R$. Since the rod is moving at constant velocity, the applied force F_a must balance the magnetic force $F_{\text{m}} = IlB$ on the rod when it is carrying the induced current I. Thus the power produced is

Equation:

$$F_a v = IlBv = \frac{Blv}{R} \cdot lBv = \frac{l^2 B^2 v^2}{R}.$$

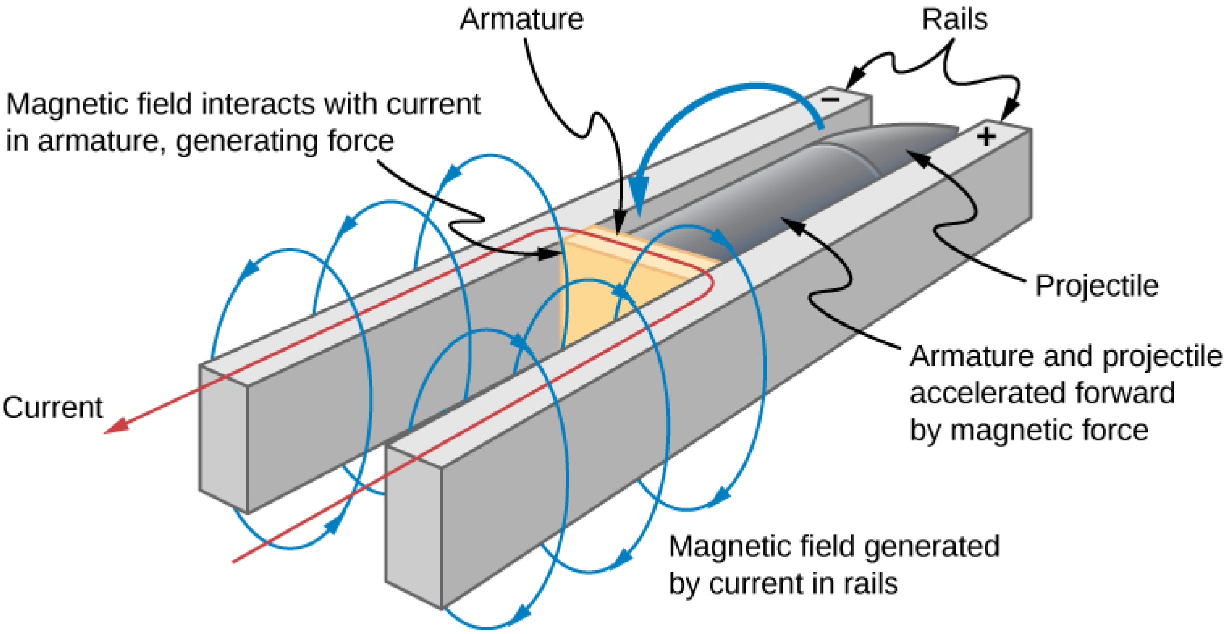
The power dissipated is

Equation:

$$P = I^2 R = \left(\frac{Blv}{R}\right)^2 R = \frac{l^2 B^2 v^2}{R}.$$

In satisfying the principle of energy conservation, the produced and dissipated powers are equal.

This principle can be seen in the operation of a rail gun. A rail gun is an electromagnetic projectile launcher that uses an apparatus similar to [link] and is shown in schematic form in [link]. The conducting rod is replaced with a projectile or weapon to be fired. So far, we’ve only heard about how motion causes an emf. In a rail gun, the optimal shutting off/ramping down of a magnetic field decreases the flux in between the rails, causing a current to flow in the rod (armature) that holds the projectile. This current through the armature experiences a magnetic force and is propelled forward. Rail guns, however, are not used widely in the military due to the high cost of production and high currents: Nearly one million amps is required to produce enough energy for a rail gun to be an effective weapon.


Armature
Rails
Magnetic field interacts with current in armature, generating force
−
+
Projectile
Armature and projectile accelerated forward by magnetic force
Current
Magnetic field generated by current in rails

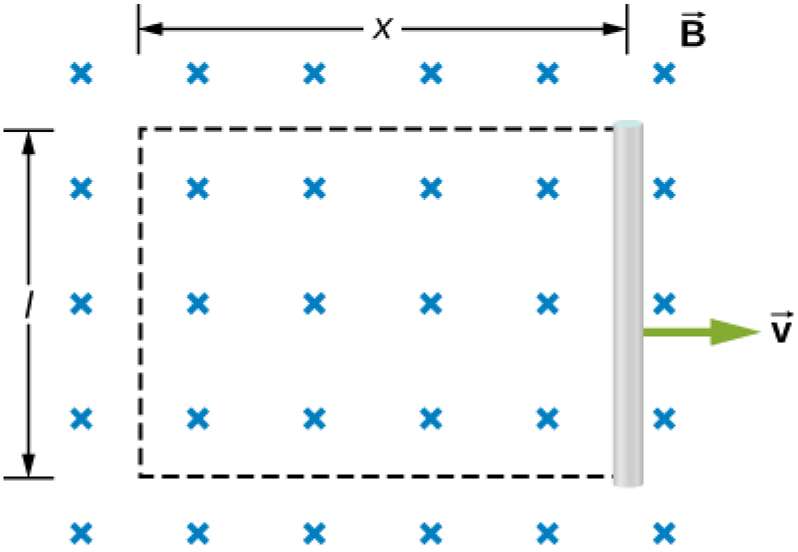
Current through two rails drives a conductive projectile forward by the magnetic force created.

We can calculate a **motionally induced emf** with Faraday’s law *even when an actual closed circuit is not present*. We simply imagine an enclosed area whose boundary includes the moving conductor, calculate Φ_{m}, and then find the emf from Faraday’s law. For example, we can let the moving rod of [link] be one side of the imaginary rectangular area represented by the dashed lines. The area of the rectangle is *lx*, so the magnetic flux through it is $\Phi_{\mathrm{m}} = Blx$. Differentiating this equation, we obtain

Equation:

$$\frac{d\Phi_{\mathrm{m}}}{dt} = Bl\frac{dx}{dt} = Blv,$$

which is identical to the potential difference between the ends of the rod that we determined earlier.



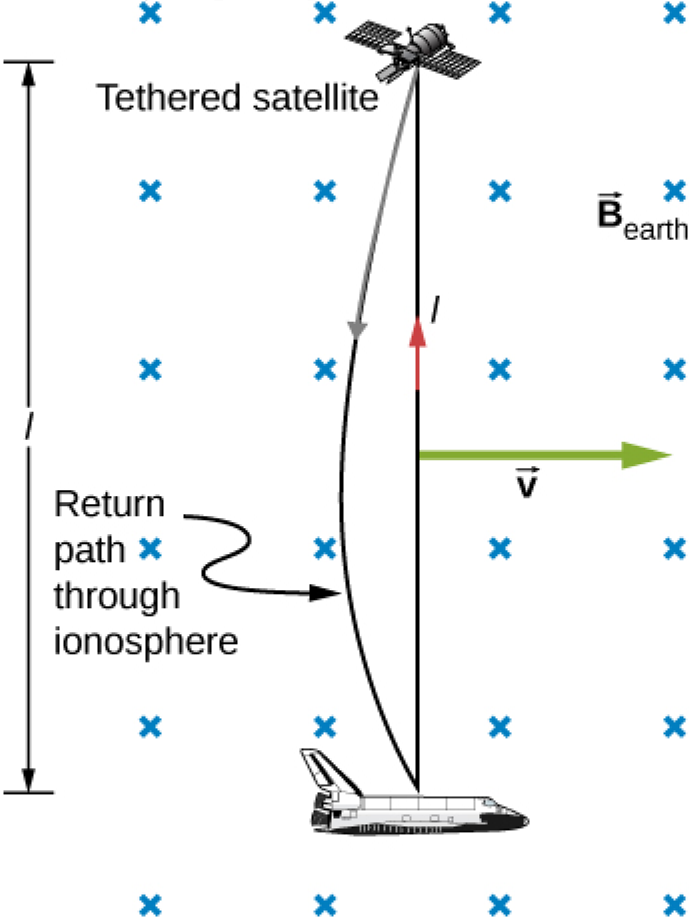
With the imaginary rectangle shown, we can use Faraday’s law to calculate the induced emf in the moving rod.

Motional emfs in Earth’s weak magnetic field are not ordinarily very large, or we would notice voltage along metal rods, such as a screwdriver, during ordinary motions. For example, a simple calculation of the motional emf of a 1.0-m rod moving at 3.0 m/s perpendicular to the Earth’s field gives

Equation:

$$\text{emf} = B\ell v = (5.0\ \times\ 10^{-5}\ \text{T})(1.0\ \text{m})(3.0\ \text{m/s}) = 150 \mu\text{V}.$$

This small value is consistent with experience. There is a spectacular exception, however. In 1992 and 1996, attempts were made with the space shuttle to create large motional emfs. The tethered satellite was to be let out on a 20-km length of wire, as shown in [link], to create a 5-kV emf by moving at orbital speed through Earth’s field. This emf could be used to convert some of the shuttle’s kinetic and potential energy into electrical energy if a complete circuit could be made. To complete the circuit, the stationary ionosphere was to supply a return path through which current could flow. (The ionosphere is the rarefied and partially ionized atmosphere at orbital altitudes. It conducts because of the ionization. The ionosphere serves the same function as the stationary rails and connecting resistor in [link], without which there would not be a complete circuit.) Drag on the current in the cable due to the magnetic force $F = I\ell B \sin \theta$ does the work that reduces the shuttle's kinetic and potential energy, and allows it to be converted into electrical energy. Both tests were unsuccessful. In the first, the cable hung up and could only be extended a couple of hundred meters; in the second, the cable broke when almost fully extended. [link] indicates feasibility in principle.



Motional emf as electrical power conversion for the space shuttle was the motivation for the tethered satellite experiment. A 5-kV emf was predicted to be induced in the 20-km tether while moving at orbital speed in Earth's magnetic field. The circuit

is completed by a return path through
the stationary ionosphere.

Example:

Calculating the Large Motional Emf of an Object in Orbit

Calculate the motional emf induced along a 20.0-km conductor moving at an orbital speed of 7.80 km/s perpendicular to Earth's $5.00 \times 10^{-5}\,\text{T}$ magnetic field.

Strategy

This is a great example of using the equation motional $\varepsilon = B\ell v$.

Solution

Entering the given values into $\varepsilon = B\ell v$ gives

Equation:

$$\begin{aligned}\varepsilon &= B\ell v \\ &= (5.00 \times 10^{-5}\,\text{T})(2.00 \times 10^{4}\,\text{m})(7.80 \times 10^{3}\,\text{m/s}) \\ &= 7.80 \times 10^{3}\,\text{V}.\end{aligned}$$

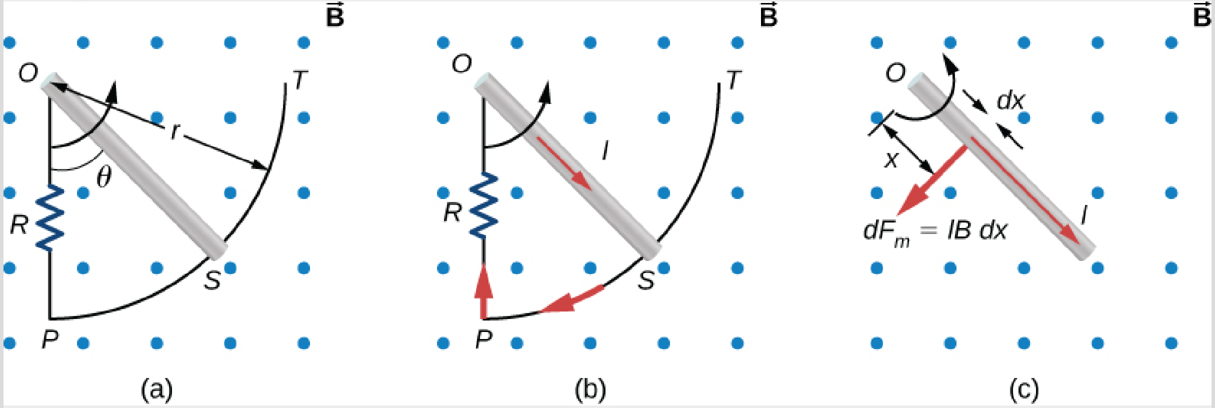
Significance

The value obtained is greater than the 5-kV measured voltage for the shuttle experiment, since the actual orbital motion of the tether is not perpendicular to Earth's field. The 7.80-kV value is the maximum emf obtained when $\theta = 90^\circ$ and so $\sin\theta = 1$.

Example:

A Metal Rod Rotating in a Magnetic Field

Part (a) of [link] shows a metal rod *OS* that is rotating in a horizontal plane around point *O*. The rod slides along a wire that forms a circular arc *PST* of radius *r*. The system is in a constant magnetic field $\vec{\mathbf{B}}$ that is directed out of the page. (a) If you rotate the rod at a constant angular velocity ω, what is the current *I* in the closed loop *OPSO*? Assume that the resistor *R* furnishes all of the resistance in the closed loop. (b) Calculate the work per unit time that you do while rotating the rod and show that it is equal to the power dissipated in the resistor.



(a) The end of a rotating metal rod slides along a circular wire in a horizontal plane. (b) The induced current in the rod. (c) The magnetic force on an infinitesimal current segment.

Strategy

The magnetic flux is the magnetic field times the area of the quarter circle or $A = r^2\theta/2$. When finding the emf through Faraday's law, all variables are constant in time but θ, with$\omega = d\theta/dt$. To calculate the work per unit time, we know this is related to the torque times the angular velocity. The torque is calculated by knowing the force on a rod and integrating it over the length of the rod.

Solution

a. From geometry, the area of the loop $OPSO$ is $A = \frac{r^2\theta}{2}$. Hence, the magnetic flux through the loop is

Equation:

$$\Phi_{\mathrm{m}} = BA = B\frac{r^2\theta}{2}.$$

Differentiating with respect to time and using $\omega = d\theta/dt$, we have

Equation:

$$\varepsilon = \left| \frac{d\Phi_{\text{m}}}{dt} \right| = \frac{Br^2\omega}{2}.$$

When divided by the resistance *R* of the loop, this yields for the magnitude of the induced current
Equation:

$$I = \frac{\varepsilon}{R} = \frac{Br^2\omega}{2R}.$$

As θ increases, so does the flux through the loop due to $\vec{\mathbf{B}}$. To counteract this increase, the magnetic field due to the induced current must be directed into the page in the region enclosed by the loop. Therefore, as part (b) of [link] illustrates, the current circulates clockwise.

b. You rotate the rod by exerting a torque on it. Since the rod rotates at constant angular velocity, this torque is equal and opposite to the torque exerted on the current in the rod by the original magnetic field. The magnetic force on the infinitesimal segment of length *dx* shown in part (c) of [link] is $dF_{\text{m}} = IBdx$, so the magnetic torque on this segment is
Equation:

$$d\tau_{\text{m}} = x \cdot dF_{\text{m}} = IBxdx.$$

The net magnetic torque on the rod is then
Equation:

$$\tau_{\text{m}} = \int_0^r d\tau_{\text{m}} = IB\int_0^r x\,dx = \frac{1}{2}IBr^2.$$

The torque τ that you exert on the rod is equal and opposite to τ_{m}, and the work that you do when the rod rotates through an angle $d\theta$ is $dW = \tau d\theta$. Hence, the work per unit time that you do on the rod is

Equation:

$$\frac{dW}{dt} = \tau\frac{d\theta}{dt} = \frac{1}{2}IBr^2\frac{d\theta}{dt} = \frac{1}{2}\left(\frac{Br^2\omega}{2R}\right)Br^2\omega = \frac{B^2r^4\omega^2}{4R},$$

where we have substituted for I. The power dissipated in the resister is $P = I^2R$, which can be written as

Equation:

$$P = \left(\frac{Br^2\omega}{2R}\right)^2 R = \frac{B^2r^4\omega^2}{4R}.$$

Therefore, we see that

Equation:

$$P = \frac{dW}{dt}.$$

Hence, the power dissipated in the resistor is equal to the work per unit time done in rotating the rod.

Significance

An alternative way of looking at the induced emf from Faraday's law is to integrate in space instead of time. The solution, however, would be the same. The motional emf is

Equation:

$$|\varepsilon| = \int Bvdl.$$

The velocity can be written as the angular velocity times the radius and the differential length written as dr. Therefore,

Equation:

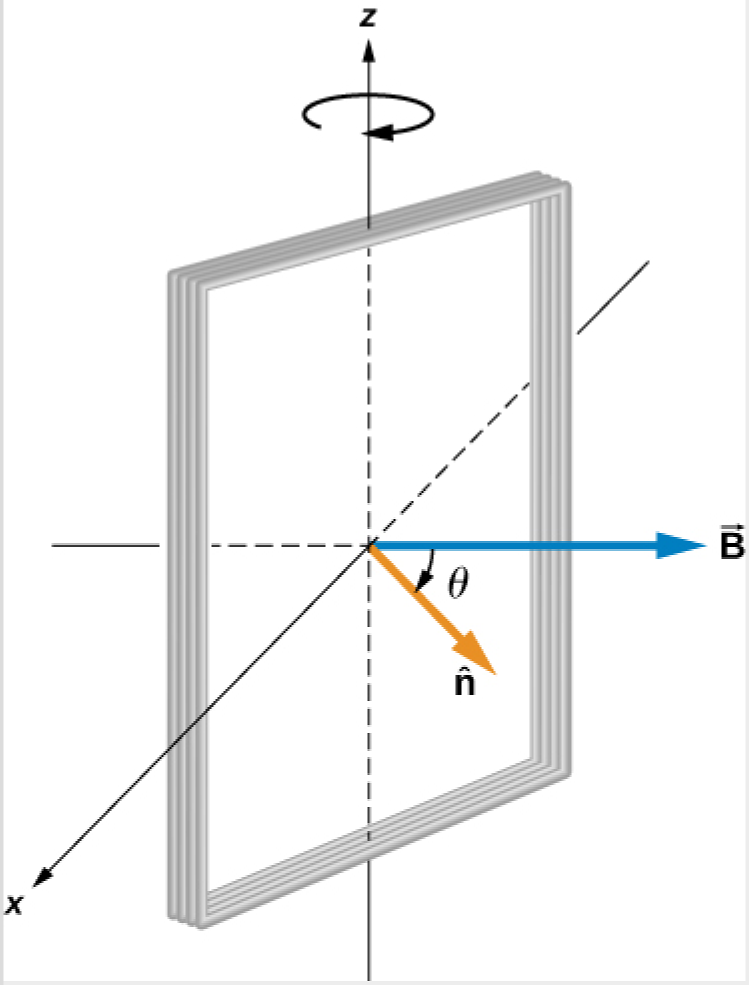
$$|\varepsilon| = B\int vdr = B\omega\int_0^l rdr = \frac{1}{2}B\omega l^2,$$

which is the same solution as before.

Example:

A Rectangular Coil Rotating in a Magnetic Field

A rectangular coil of area *A* and *N* turns is placed in a uniform magnetic field $\vec{\mathbf{B}} = B\hat{\mathbf{j}}$, as shown in [link]. The coil is rotated about the *z*-axis through its center at a constant angular velocity ω. Obtain an expression for the induced emf in the coil.



A rectangular coil rotating in a uniform magnetic field.

Strategy
According to the diagram, the angle between the perpendicular to the surface ($\hat{n}$) and the magnetic field $(\vec{\mathbf{B}})$ is θ. The dot product of $\vec{\mathbf{B}} \cdot \hat{\mathbf{n}}$ simplifies to only the $\cos\theta$ component of the magnetic field, namely where the magnetic field projects onto the unit area vector $\hat{n}$. The magnitude of the magnetic field and the area of the loop are fixed over time, which makes the integration simplify quickly. The induced emf is written out using Faraday’s law.

Solution

When the coil is in a position such that its normal vector $\hat{\mathbf{n}}$ makes an angle θ with the magnetic field $\vec{\mathbf{B}}$, the magnetic flux through a single turn of the coil is

Equation:

$$\Phi_{\mathrm{m}} = \int_S \vec{\mathbf{B}} \cdot \hat{\mathbf{n}} dA = BA\cos\theta.$$

From Faraday's law, the emf induced in the coil is

Equation:

$$\varepsilon = -N\frac{d\Phi_{\mathrm{m}}}{dt} = NBA\sin\theta\frac{d\theta}{dt}.$$

The constant angular velocity is $\omega = d\theta/dt$. The angle θ represents the time evolution of the angular velocity or ωt. This is changes the function to time space rather than θ. The induced emf therefore varies sinusoidally with time according to

Equation:

$$\varepsilon = \varepsilon_0 \sin\omega t,$$

where $\varepsilon_0 = NBA\omega$.

Significance

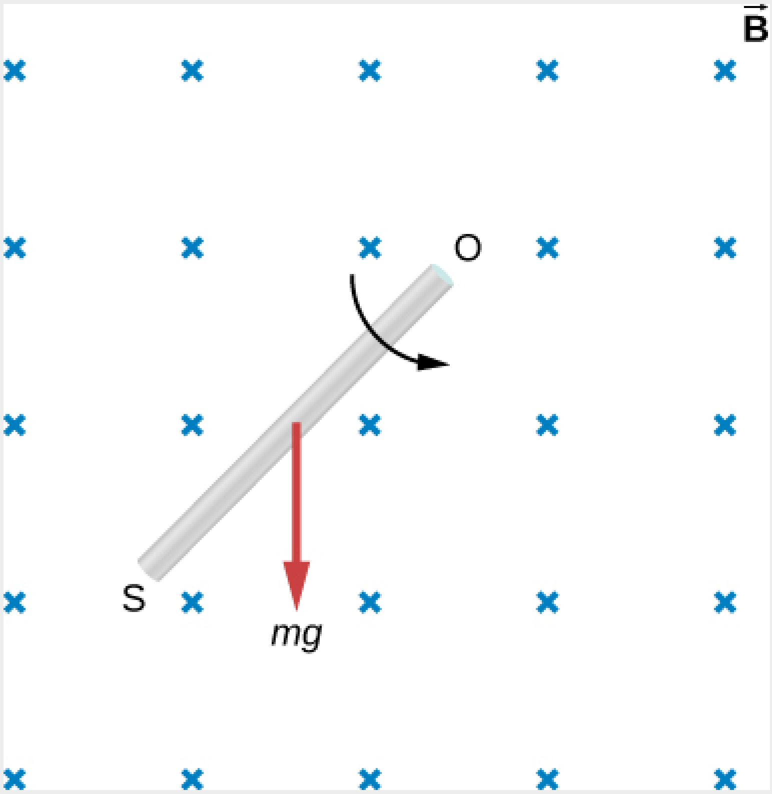
If the magnetic field strength or area of the loop were also changing over time, these variables wouldn't be able to be pulled out of the time derivative to simply the solution as shown. This example is the basis for an electric generator, as we will give a full discussion in Applications of Newton's Law.

Note:

Exercise:

Problem:

Check Your Understanding Shown below is a rod of length l that is rotated counterclockwise around the axis through O by the torque due to $m\vec{\mathbf{g}}$. Assuming that the rod is in a uniform magnetic field $\vec{\mathbf{B}}$, what is the emf induced between the ends of the rod when its angular velocity is ω? Which end of the rod is at a higher potential?



Solution:

$\varepsilon = Bl^2\omega/2$, with O at a higher potential than S

Note:
Exercise:

Problem:

Check Your Understanding A rod of length 10 cm moves at a speed of 10 m/s perpendicularly through a 1.5-T magnetic field. What is the potential difference between the ends of the rod?

Solution:

1.5 V

Summary

- The relationship between an induced emf ε in a wire moving at a constant speed *v* through a magnetic field *B* is given by $\varepsilon = Blv$.
- An induced emf from Faraday’s law is created from a motional emf that opposes the change in flux.

Conceptual Questions

Exercise:

Problem:

A bar magnet falls under the influence of gravity along the axis of a long copper tube. If air resistance is negligible, will there be a force to oppose the descent of the magnet? If so, will the magnet reach a terminal velocity?

Exercise:

Problem:

Around the geographic North Pole (or magnetic South Pole), Earth’s magnetic field is almost vertical. If an airplane is flying northward in this region, which side of the wing is positively charged and which is negatively charged?

Solution:

Positive charges on the wings would be to the west, or to the left of the pilot while negative charges would be pulled east or to the right of the pilot. Thus, the left hand tips of the wings would be positive and the right hand tips would be negative.

Exercise:

Problem:

A wire loop moves translationally (no rotation) in a uniform magnetic field. Is there an emf induced in the loop?

Problems

Exercise:

Problem:

An automobile with a radio antenna 1.0 m long travels at 100.0 km/h in a location where the Earth's horizontal magnetic field is $5.5 \times 10^{-5}\,\mathrm{T}$. What is the maximum possible emf induced in the antenna due to this motion?

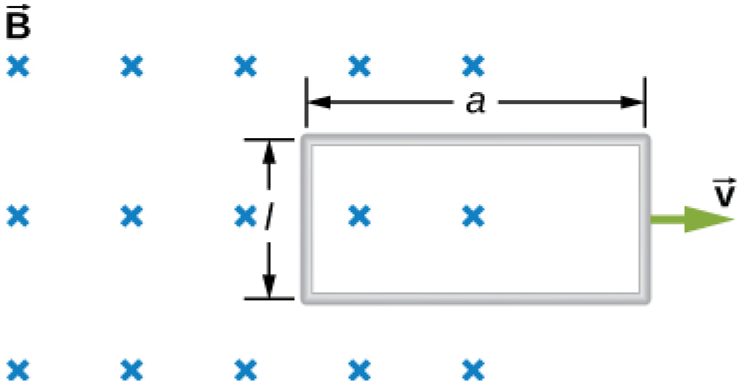
Solution:

0.0015 V

Exercise:

Problem:

The rectangular loop of *N* turns shown below moves to the right with a constant velocity $\vec{\mathbf{v}}$ while leaving the poles of a large electromagnet. (a) Assuming that the magnetic field is uniform between the pole faces and negligible elsewhere, determine the induced emf in the loop. (b) What is the source of work that produces this emf?



Exercise:

Problem:

Suppose the magnetic field of the preceding problem oscillates with time according to $B = B_0 \sin \omega t$. What then is the emf induced in the loop when its trailing side is a distance *d* from the right edge of the magnetic field region?

Solution:

$\varepsilon = -B_0 l d \omega \cos\left(\Omega t\right)\mathrm{ld} + B_0 \mathrm{sin}(\Omega t)\mathrm{lv}$

Exercise:

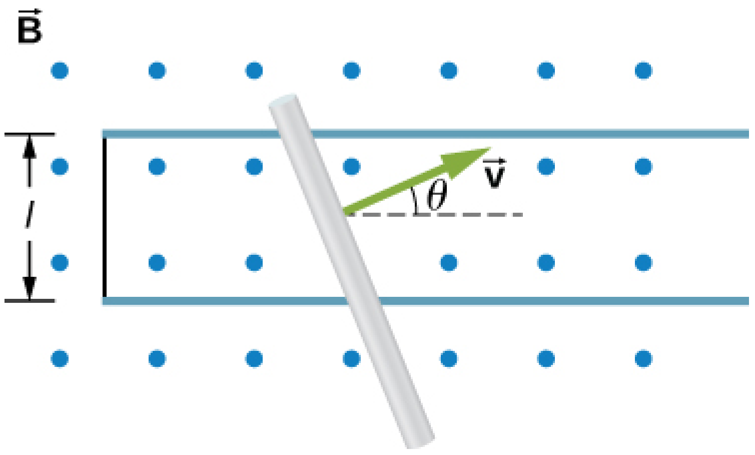
Problem:

A coil of 1000 turns encloses an area of $25\ \mathrm{cm}^2$. It is rotated in 0.010 s from a position where its plane is perpendicular to Earth's magnetic field to one where its plane is parallel to the field. If the strength of the field is $6.0\ \times\ 10^{-5}\ \mathrm{T}$, what is the average emf induced in the coil?

Exercise:

Problem:

In the circuit shown in the accompanying figure, the rod slides along the conducting rails at a constant velocity $\vec{\mathbf{v}}$. The velocity is in the same plane as the rails and directed at an angle θ to them. A uniform magnetic field $\vec{\mathbf{B}}$ is directed out of the page. What is the emf induced in the rod?



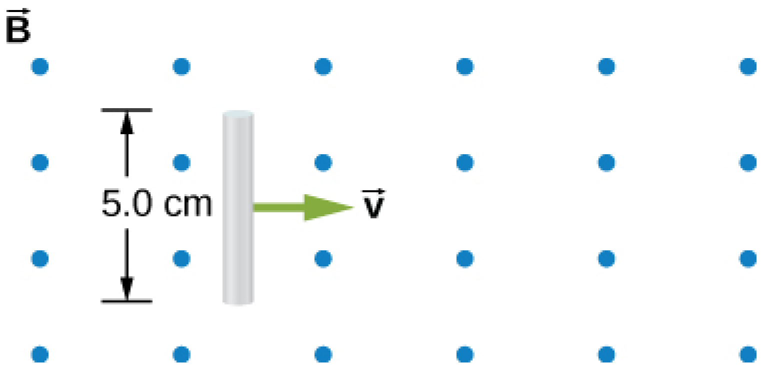
Solution:

$\varepsilon = Blv\cos\theta$

Exercise:

Problem:

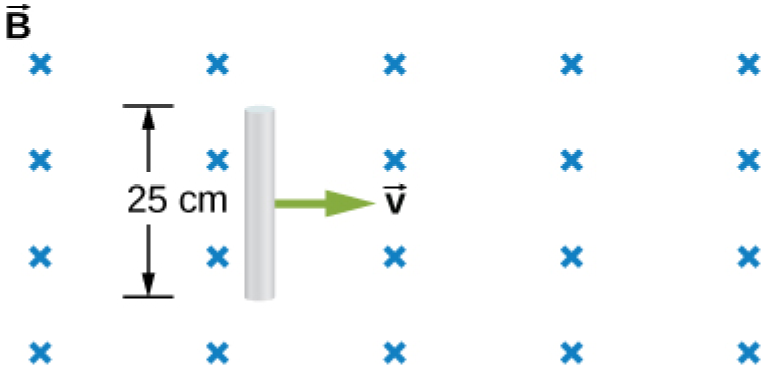
The rod shown in the accompanying figure is moving through a uniform magnetic field of strength $B = 0.50\,\mathrm{T}$ with a constant velocity of magnitude $v = 8.0\,\mathrm{m/s}$. What is the potential difference between the ends of the rod? Which end of the rod is at a higher potential?



Exercise:

Problem:

A 25-cm rod moves at 5.0 m/s in a plane perpendicular to a magnetic field of strength 0.25 T. The rod, velocity vector, and magnetic field vector are mutually perpendicular, as indicated in the accompanying figure. Calculate (a) the magnetic force on an electron in the rod, (b) the electric field in the rod, and (c) the potential difference between the ends of the rod. (d) What is the speed of the rod if the potential difference is 1.0 V?



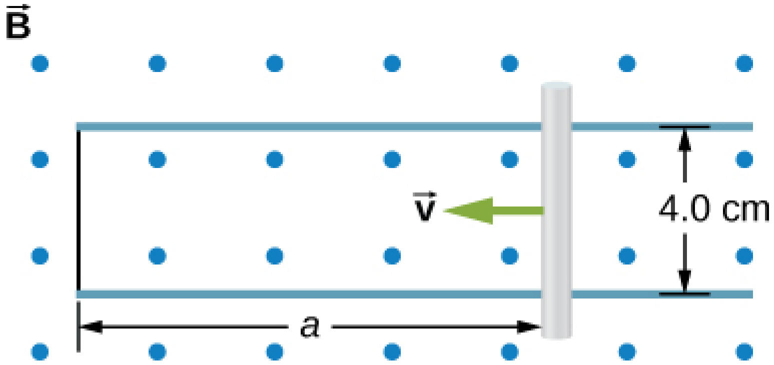
Solution:

a. $2\ \times\ 10^{-19}T$; b. 1.25 V/m; c. 0.3125 V; d. 16 m/s

Exercise:

Problem:

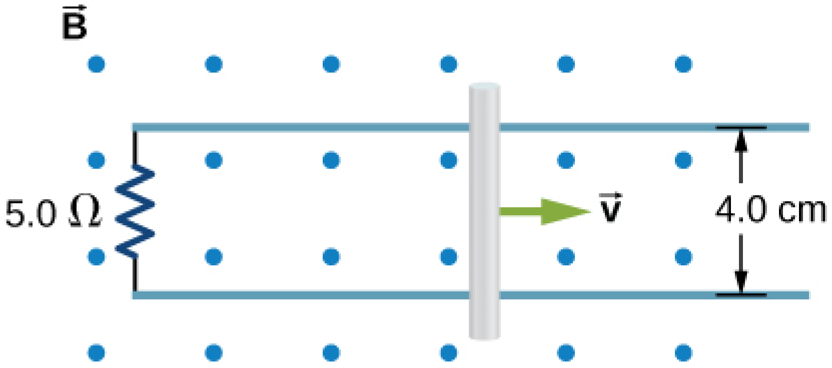
In the accompanying figure, the rails, connecting end piece, and rod all have a resistance per unit length of $2.0\,\Omega/\text{cm}$. The rod moves to the left at $v = 3.0\,\text{m/s}$. If $B = 0.75\,\text{T}$ everywhere in the region, what is the current in the circuit (a) when $a = 8.0\,\text{cm}$? (b) when $a = 5.0\,\text{cm}$? Specify also the sense of the current flow.



Exercise:

Problem:

The rod shown below moves to the right on essentially zero-resistance rails at a speed of $v = 3.0\,\mathrm{m/s}$. If $B = 0.75\,\mathrm{T}$ everywhere in the region, what is the current through the 5.0-Ω resistor? Does the current circulate clockwise or counterclockwise?



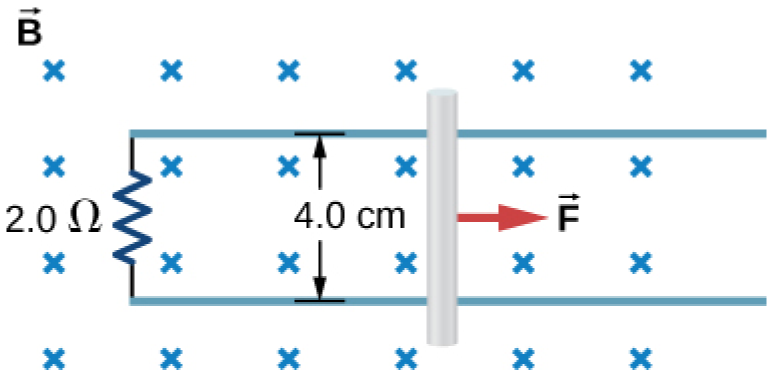
Solution:

0.018 A, CW as seen in the diagram

Exercise:

Problem:

Shown below is a conducting rod that slides along metal rails. The apparatus is in a uniform magnetic field of strength 0.25 T, which is directly into the page. The rod is pulled to the right at a constant speed of 5.0 m/s by a force $\vec{\mathbf{F}}$. The only significant resistance in the circuit comes from the 2.0-Ω resistor shown. (a) What is the emf induced in the circuit? (b) What is the induced current? Does it circulate clockwise or counter clockwise? (c) What is the magnitude of $\vec{\mathbf{F}}$? (d) What are the power output of $\vec{\mathbf{F}}$ and the power dissipated in the resistor?



Glossary

motionally induced emf
: voltage produced by the movement of a conducting wire in a magnetic field

Induced Electric Fields

By the end of this section, you will be able to:

- Connect the relationship between an induced emf from Faraday’s law to an electric field, thereby showing that a changing magnetic flux creates an electric field
- Solve for the electric field based on a changing magnetic flux in time

The fact that emfs are induced in circuits implies that work is being done on the conduction electrons in the wires. What can possibly be the source of this work? We know that it’s neither a battery nor a magnetic field, for a battery does not have to be present in a circuit where current is induced, and magnetic fields never do work on moving charges. The answer is that the source of the work is an electric field $\vec{\mathbf{E}}$ that is induced in the wires. The work done by $\vec{\mathbf{E}}$ in moving a unit charge completely around a circuit is the induced emf ε; that is,

Equation:

$$\varepsilon = \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}},$$

where $\oint$ represents the line integral around the circuit. Faraday’s law can be written in terms of the **induced electric field** as

Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = -\frac{d\Phi_{\mathrm{m}}}{dt}.$$

There is an important distinction between the electric field induced by a changing magnetic field and the electrostatic field produced by a fixed charge distribution. Specifically, the induced electric field is nonconservative because it does net work in moving a charge over a closed path, whereas the electrostatic field is conservative and does no net work over a closed path. Hence, electric potential can be associated with the electrostatic field, but not with the induced field. The following equations represent the distinction between the two types of electric field:

Equation:

$$\begin{aligned}\oint \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}} &\neq 0 \text{ (induced);}\\ \oint \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}} &= 0 \text{ (electrostatic).}\end{aligned}$$

Our results can be summarized by combining these equations:

Note:

Equation:

$$\varepsilon = \oint \vec{\mathbf{E}}\cdot d\vec{\mathbf{l}} = -\frac{d\Phi_{\text{m}}}{dt}.$$

Example:

Induced Electric Field in a Circular Coil

What is the induced electric field in the circular coil of [link] (and [link]) at the three times indicated?

Strategy

Using cylindrical symmetry, the electric field integral simplifies into the electric field times the circumference of a circle. Since we already know the induced emf, we can connect these two expressions by Faraday's law to solve for the induced electric field.

Solution

The induced electric field in the coil is constant in magnitude over the cylindrical surface, similar to how Ampere's law problems with cylinders are solved. Since $\vec{\mathbf{E}}$ is tangent to the coil,

Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = \oint E dl = 2\pi r E.$$

When combined with [link], this gives

Equation:

$$E = \frac{\varepsilon}{2\pi r}.$$

The direction of ε is counterclockwise, and $\vec{\mathbf{E}}$ circulates in the same direction around the coil. The values of E are

Equation:

$$\begin{aligned} E(t_1) &= \frac{6.0\,\text{V}}{2\pi\,(0.50\,\text{m})} = 1.9\,\text{V/m}; \\ E(t_2) &= \frac{4.7\,\text{V}}{2\pi\,(0.50\,\text{m})} = 1.5\,\text{V/m}; \\ E(t_3) &= \frac{0.040\,\text{V}}{2\pi\,(0.50\,\text{m})} = 0.013\,\text{V/m}. \end{aligned}$$

Significance

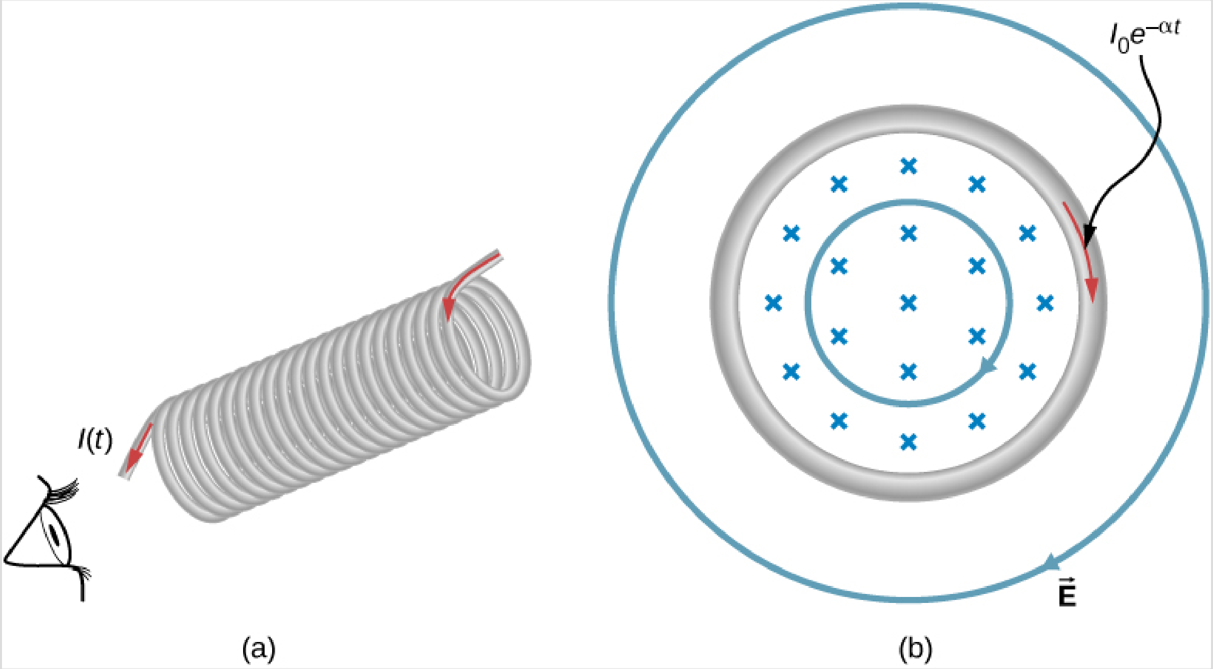
When the magnetic flux through a circuit changes, a nonconservative electric field is induced, which drives current through the circuit. But what happens if $dB/dt \neq 0$ in free space where there isn't a conducting path? The answer is that this case can be treated *as if a conducting path were present*; that is, nonconservative electric fields are induced wherever $dB/dt \neq 0$, whether or not there is a conducting path present.

These nonconservative electric fields always satisfy [link]. For example, if the circular coil of [link] were removed, an electric field *in free space* at $r = 0.50\,\text{m}$ would still be directed counterclockwise, and its magnitude would still be 1.9 V/m at $t = 0$, 1.5 V/m at $t = 5.0 \times 10^{-2}$ s, etc. The existence of induced electric fields is certainly *not* restricted to wires in circuits.

Example:

Electric Field Induced by the Changing Magnetic Field of a Solenoid

Part (a) of [link] shows a long solenoid with radius R and n turns per unit length; its current decreases with time according to $I = I_0 e^{-\alpha t}$. What is the magnitude of the induced electric field at a point a distance r from the central axis of the solenoid (a) when $r > R$ and (b) when $r < R$ [see part (b) of [link]]. (c) What is the direction of the induced field at both locations? Assume that the infinite-solenoid approximation is valid throughout the regions of interest.



(a) The current in a long solenoid is decreasing exponentially. (b) A cross-sectional view of the solenoid from its left end. The cross-section shown is near the middle of the solenoid. An electric field is induced both inside and outside the solenoid.

Strategy

Using the formula for the magnetic field inside an infinite solenoid and Faraday’s law, we calculate the induced emf. Since we have cylindrical symmetry, the electric field integral reduces to the electric field times the circumference of the integration path. Then we solve for the electric field.

Solution

a. The magnetic field is confined to the interior of the solenoid where

Equation:

$$B = \mu_0 n I = \mu_0 n I_0 e^{-\alpha t}.$$

Thus, the magnetic flux through a circular path whose radius *r* is greater than *R*, the solenoid radius, is

Equation:

$$\Phi_{\mathrm{m}} = BA = \mu_0 n I_0 \pi R^2 e^{-\alpha t}.$$

The induced field $\vec{\mathbf{E}}$ is tangent to this path, and because of the cylindrical symmetry of the system, its magnitude is constant on the path. Hence, we have

Equation:

$$\begin{aligned}\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} &= \tfrac{d\Phi_{\mathrm{m}}}{dt}\ , \\ E(2\pi r) &= \tfrac{d}{dt}\left(\mu_0 n I_0 \pi R^2 e^{-\alpha t}\right)\ = \alpha\mu_0 n I_0 \pi R^2 e^{-\alpha t}, \\ E &= \tfrac{\alpha\mu_0 n I_0 R^2}{2r} e^{-\alpha t}\ \ (r > R).\end{aligned}$$

b. For a path of radius *r* inside the solenoid, $\Phi_{\mathrm{m}} = B\pi r^2$, so

Equation:

$$E(2\pi r) = \frac{d}{dt}\left(\mu_0 n I_0 \pi r^2 e^{-\alpha t}\right)\ = \alpha\mu_0 n I_0 \pi r^2 e^{-\alpha t},$$

and the induced field is

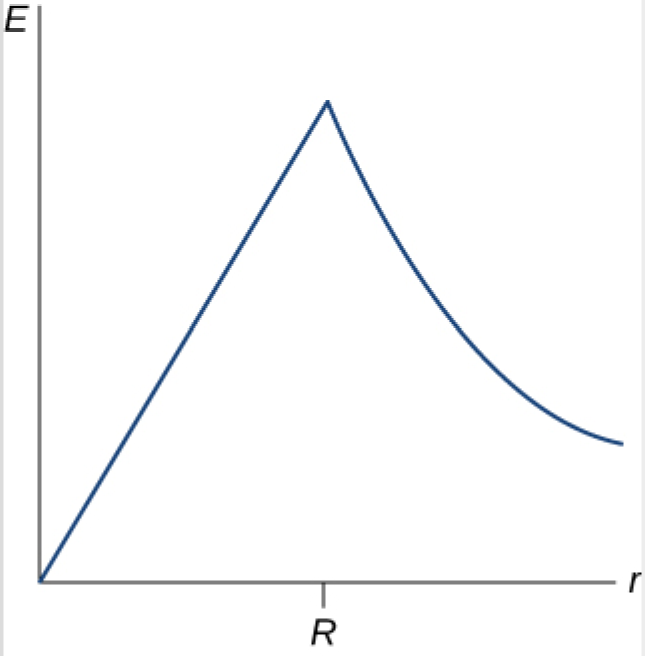
Equation:

$$E = \frac{\alpha\mu_0 n I_0 r}{2} e^{-\alpha t}\ (r < R).$$

c. The magnetic field points into the page as shown in part (b) and is decreasing. If either of the circular paths were occupied by conducting rings, the currents induced in them would circulate as shown, in conformity with Lenz's law. The induced electric field must be so directed as well.

Significance

In part (b), note that $\vec{\mathbf{E}}$ increases with r inside and decreases as $1/r$ outside the solenoid, as shown in [link].



The electric field vs. distance r. When $r < R$, the electric field rises linearly, whereas when $r > R$, the electric field falls of proportional to $1/r$.

Note:
Exercise:

Problem:

Check Your Understanding Suppose that the coil of [link] is a square rather than circular. Can [link] be used to calculate (a) the induced emf and (b) the induced electric field?

Solution:

a. yes; b. Yes; however there is a lack of symmetry between the electric field and coil, making $\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}}$ a more complicated relationship that can’t be simplified as shown in the example.

Note:
Exercise:

Problem:

Check Your Understanding What is the magnitude of the induced electric field in [link] at $t = 0$ if $r = 6.0\ \text{cm}$, $R = 2.0\ \text{cm}$, $n = 2000$ turns per meter, $I_0 = 2.0\ \text{A}$, and $\alpha = 200\ \text{s}^{-1}$?

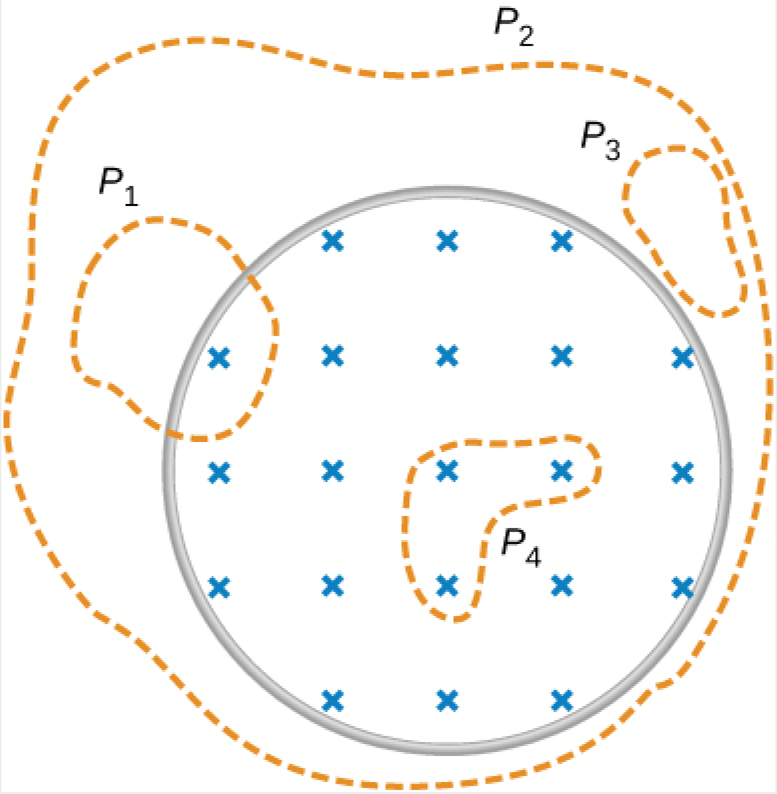
Solution:

$3.4\ \times\ 10^{-3}\ \text{V/m}$

Note:
Exercise:

Problem:

Check Your Understanding The magnetic field shown below is confined to the cylindrical region shown and is changing with time. Identify those paths for which $\varepsilon = \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} \neq 0$.



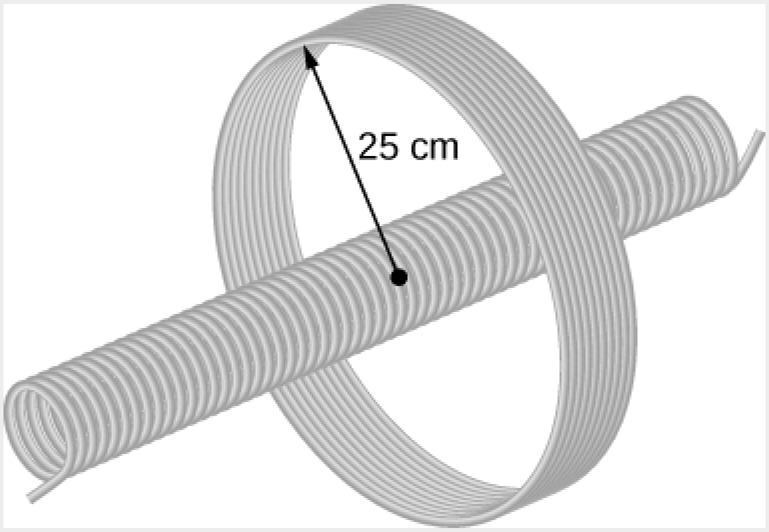
Solution:

P_1, P_2, P_4

Note:
Exercise:

Problem:

Check Your Understanding A long solenoid of cross-sectional area $5.0\ \text{cm}^2$ is wound with 25 turns of wire per centimeter. It is placed in the middle of a closely wrapped coil of 10 turns and radius 25 cm, as shown below. (a) What is the emf induced in the coil when the current through the solenoid is decreasing at a rate $dI/dt = -0.20\ \text{A/s}$? (b) What is the electric field induced in the coil?



Solution:

a. $3.1\ \times\ 10^{-6}\text{V}$; b. $2.0\ \times\ 10^{-7}\ \text{V/m}$

Summary

- A changing magnetic flux induces an electric field.
- Both the changing magnetic flux and the induced electric field are related to the induced emf from Faraday’s law.

Conceptual Questions

Exercise:

Problem:

Is the work required to accelerate a rod from rest to a speed *v* in a magnetic field greater than the final kinetic energy of the rod? Why?

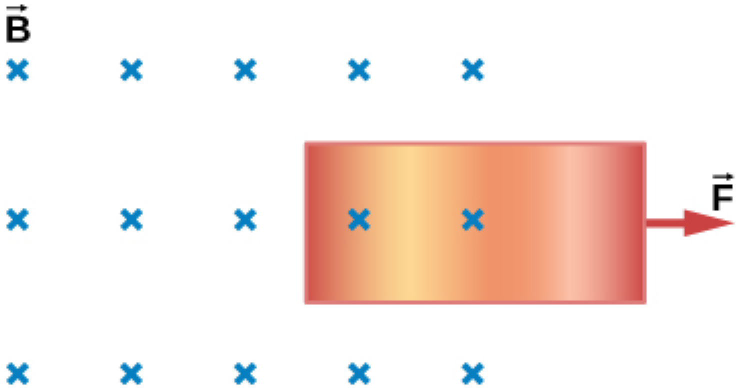
Solution:

The work is greater than the kinetic energy because it takes energy to counteract the induced emf.

Exercise:

Problem:

The copper sheet shown below is partially in a magnetic field. When it is pulled to the right, a resisting force pulls it to the left. Explain. What happen if the sheet is pushed to the left?



Problems

Exercise:

Problem:

Calculate the induced electric field in a 50-turn coil with a diameter of 15 cm that is placed in a spatially uniform magnetic field of magnitude 0.50 T so that the face of the coil and the magnetic field are perpendicular. This magnetic field is reduced to zero in 0.10 seconds. Assume that the magnetic field is cylindrically symmetric with respect to the central axis of the coil.

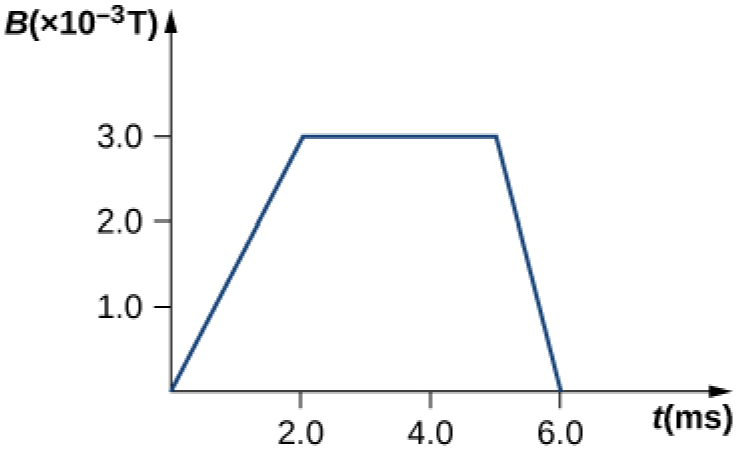
Solution:

9.375 V/m

Exercise:

Problem:

The magnetic field through a circular loop of radius 10.0 cm varies with time as shown in the accompanying figure. The field is perpendicular to the loop. Assuming cylindrical symmetry with respect to the central axis of the loop, plot the induced electric field in the loop as a function of time.



Exercise:

Problem:

The current *I* through a long solenoid with *n* turns per meter and radius *R* is changing with time as given by *dI*/*dt*. Calculate the induced electric field as a function of distance *r* from the central axis of the solenoid.

Solution:

Inside, $B = \mu_0 nI$, $\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = \pi r^2 \ \mu_0 n \frac{dI}{dt}$, so, $E = \frac{\mu_0 n r}{2} \cdot \frac{dI}{dt}$ (inside). Outside, $E\left(2\pi r\right) = \pi R^2 \mu_0 n \frac{dI}{dt}$, so, $E = \frac{\mu_0 n R^2}{2r} \cdot \frac{dI}{dt}$ (outside)

Exercise:

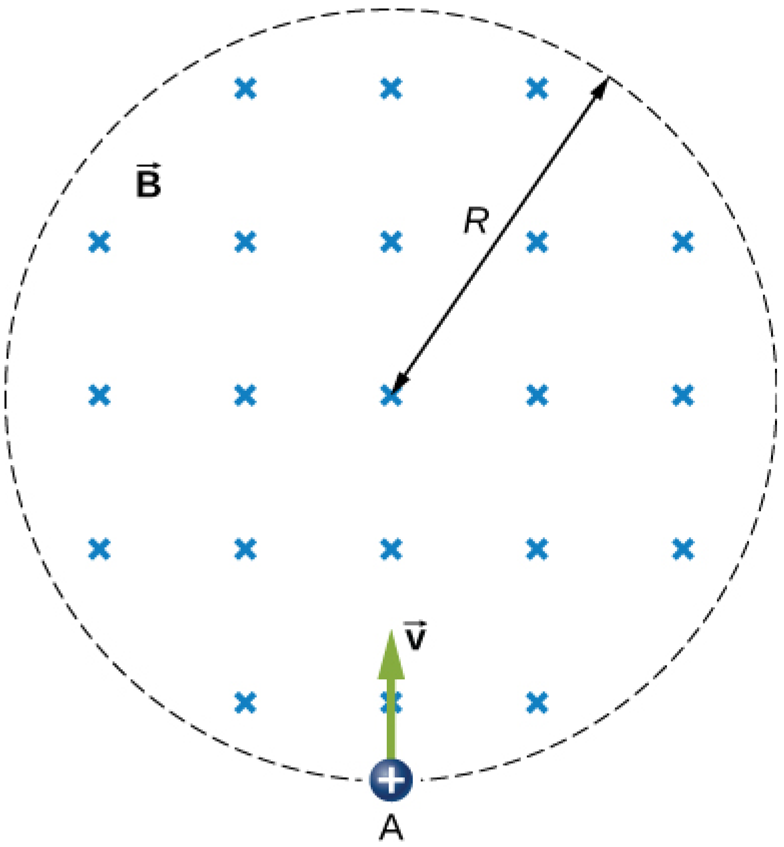
Problem:

Calculate the electric field induced both inside and outside the solenoid of the preceding problem if $I = I_0 \sin \omega t$.

Exercise:

Problem:

Over a region of radius *R*, there is a spatially uniform magnetic field $\vec{\mathbf{B}}$. (See below.) At $t = 0$, $B = 1.0 \text{ T}$, after which it decreases at a constant rate to zero in 30 s. (a) What is the electric field in the regions where $r \le R$ and $r \ge R$ during that 30-s interval? (b) Assume that $R = 10.0 \text{ cm}$. How much work is done by the electric field on a proton that is carried once clock wise around a circular path of radius 5.0 cm? (c) How much work is done by the electric field on a proton that is carried once counterclockwise around a circular path of any radius $r \ge R$? (d) At the instant when $B = 0.50 \text{ T}$, a proton enters the magnetic field at *A*, moving a velocity $\vec{\mathbf{v}}$ $v = 5.0 \ \times \ 10^6 \text{ m/s}$ as shown. What are the electric and magnetic forces on the proton at that instant?



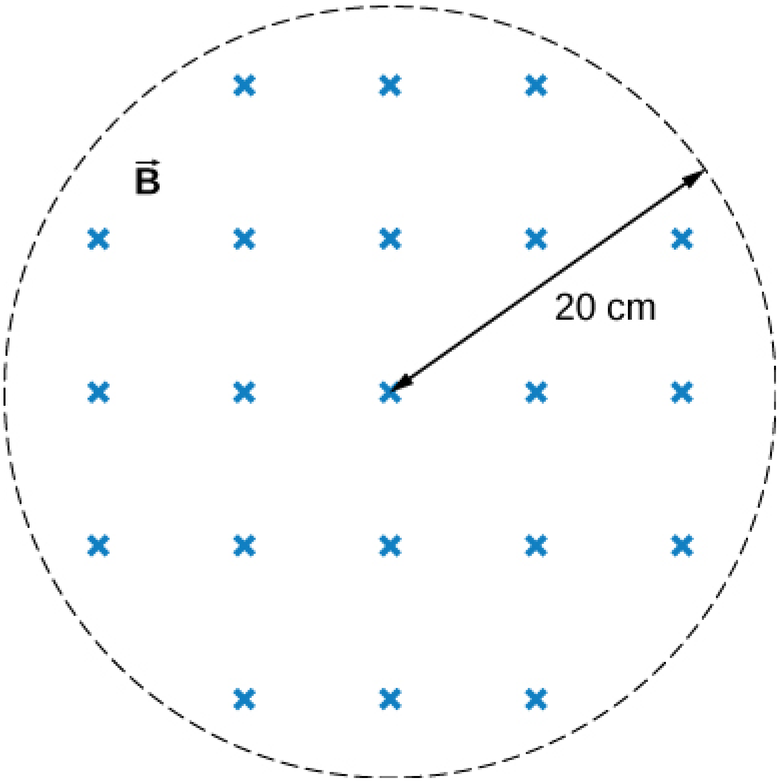
Solution:

a. $E_{\text{inside}} = \frac{r}{2}\frac{dB}{dt}$, $E_{\text{outside}} = \frac{r^2}{2R}\frac{dB}{dt}$; b. $W = 4.19 \times 10^{-23}$ J; c. 0 J; d. $F_{\text{mag}} = 4 \times 10^{-13}$ N, $F_{\text{elec}} = 2.7 \times 10^{-22}$ N

Exercise:

Problem:

The magnetic field at all points within the cylindrical region whose cross-section is indicated in the accompanying figure starts at 1.0 T and decreases uniformly to zero in 20 s. What is the electric field (both magnitude and direction) as a function of *r*, the distance from the geometric center of the region?



Exercise:

Problem:

The current in a long solenoid with 20 turns per centimeter of radius 3 cm is varied with time at a rate of 2 A/s. A circular loop of wire of radius 5 cm and resistance $2\;\Omega$ surrounds the solenoid. Find the electrical current induced in the loop.

Solution:

$7.1\;\mu\text{A}$

Exercise:

Problem:

The current in a long solenoid of radius 3 cm and 20 turns/cm is varied with time at a rate of 2 A/s. Find the electric field at a distance of 4 cm from the center of the solenoid.

Glossary

induced electric field
: created based on the changing magnetic flux with time

Eddy Currents

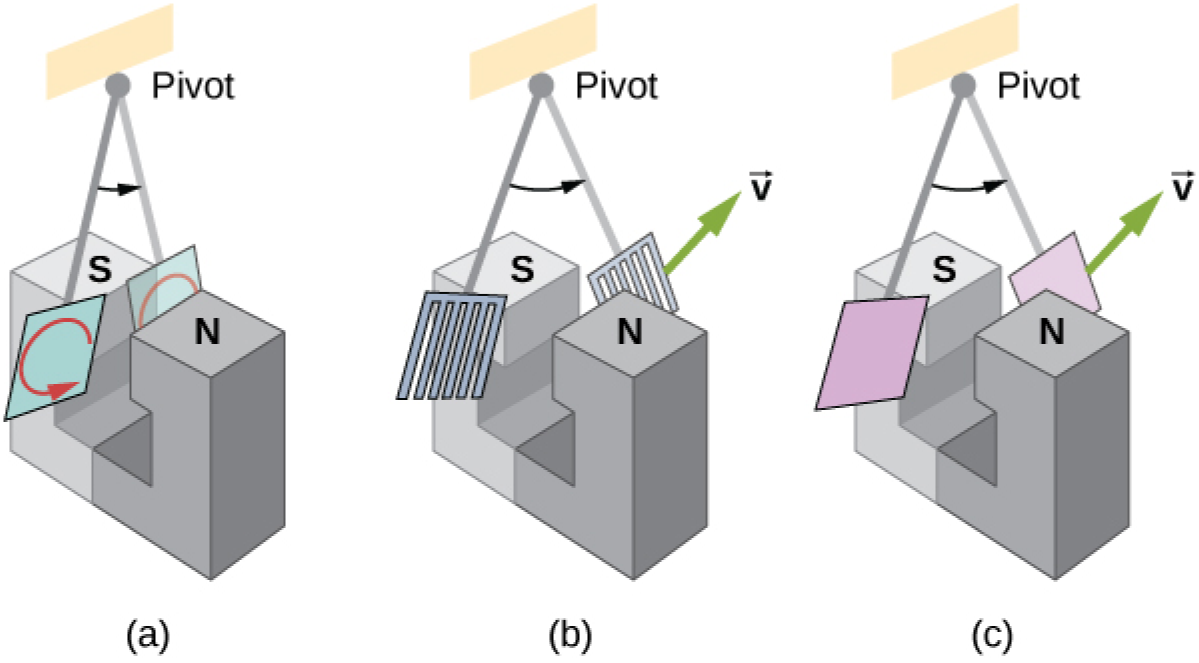
By the end of this section, you will be able to:

- Explain how eddy currents are created in metals
- Describe situations where eddy currents are beneficial and where they are not helpful

As discussed two sections earlier, a motional emf is induced when a conductor moves in a magnetic field or when a magnetic field moves relative to a conductor. If motional emf can cause a current in the conductor, we refer to that current as an **eddy current**.

Magnetic Damping

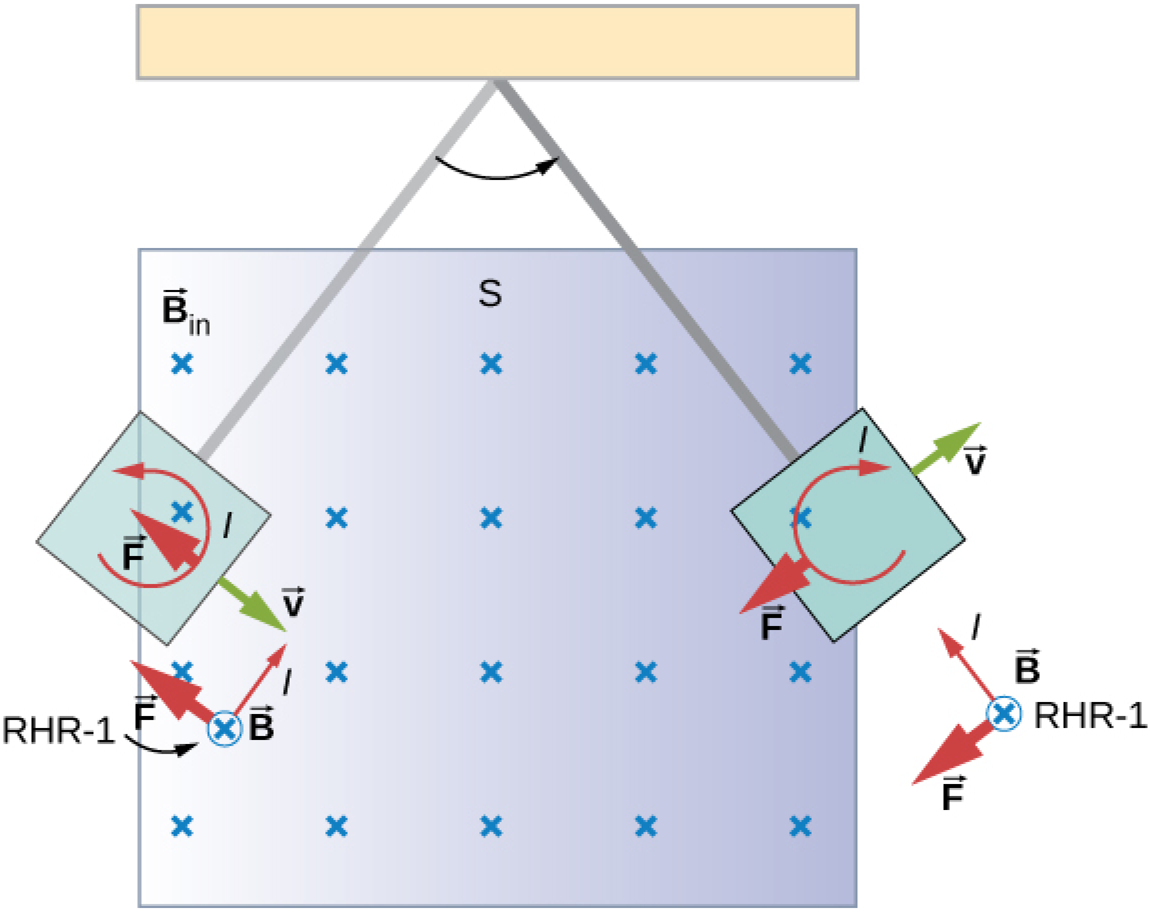
Eddy currents can produce significant drag, called **magnetic damping**, on the motion involved. Consider the apparatus shown in [link], which swings a pendulum bob between the poles of a strong magnet. (This is another favorite physics demonstration.) If the bob is metal, significant drag acts on the bob as it enters and leaves the field, quickly damping the motion. If, however, the bob is a slotted metal plate, as shown in part (b) of the figure, the magnet produces a much smaller effect. There is no discernible effect on a bob made of an insulator. Why does drag occur in both directions, and are there any uses for magnetic drag?



A common physics demonstration device for exploring eddy currents and magnetic damping. (a) The motion of a metal pendulum bob swinging between the poles of a magnet is quickly damped by the action of eddy currents. (b) There is little effect on the motion of a slotted metal bob, implying that eddy currents are made less effective. (c) There is also no magnetic damping on a nonconducting bob, since the eddy currents are extremely small.

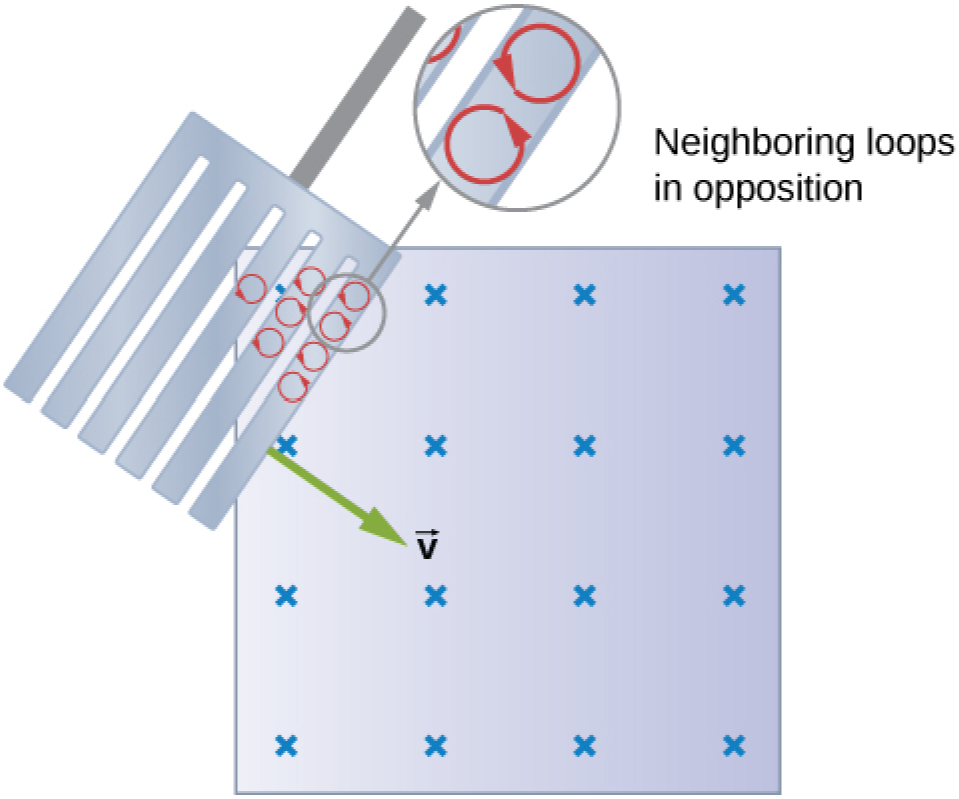
[link] shows what happens to the metal plate as it enters and leaves the magnetic field. In both cases, it experiences a force opposing its motion. As it enters from the left, flux increases, setting up an eddy current (Faraday's law) in the counterclockwise direction (Lenz's law), as shown. Only the right-hand side of the current loop is in the field, so an unopposed force acts on it to the left (RHR-1). When the metal plate is completely inside the field, there is no eddy current if the field is uniform, since the flux remains constant in this region. But when the plate leaves the field on the right, flux decreases, causing an eddy current in the clockwise direction that, again, experiences a force to the left, further slowing the motion. A similar analysis of what happens when the plate swings from the right toward the

left shows that its motion is also damped when entering and leaving the field.



A more detailed look at the conducting plate passing between the poles of a magnet. As it enters and leaves the field, the change in flux produces an eddy current. Magnetic force on the current loop opposes the motion. There is no current and no magnetic drag when the plate is completely inside the uniform field.

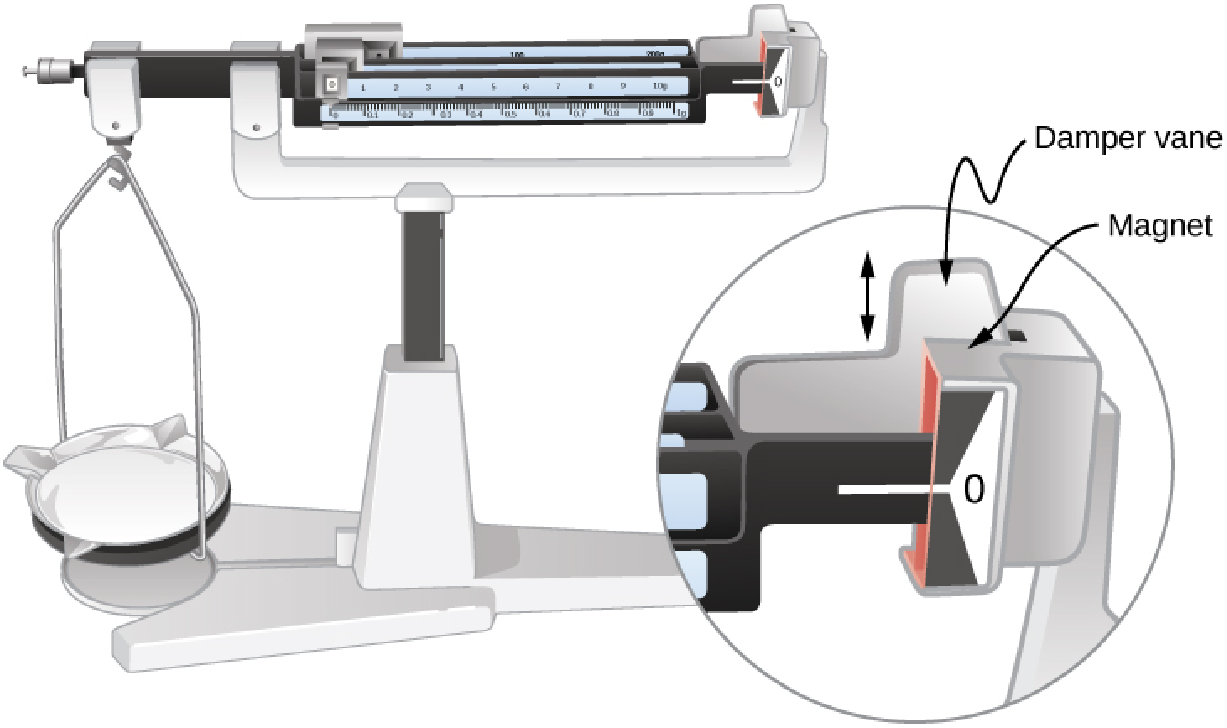
When a slotted metal plate enters the field ([link]), an emf is induced by the change in flux, but it is less effective because the slots limit the size of the current loops. Moreover, adjacent loops have currents in opposite directions, and their effects cancel. When an insulating material is used, the eddy current is extremely small, so magnetic damping on insulators is negligible. If eddy currents are to be avoided in conductors, then they must be slotted or constructed of thin layers of conducting material separated by insulating sheets.



Eddy currents induced in a slotted metal plate entering a magnetic field form small loops, and the forces on them tend to cancel, thereby making magnetic drag almost zero.

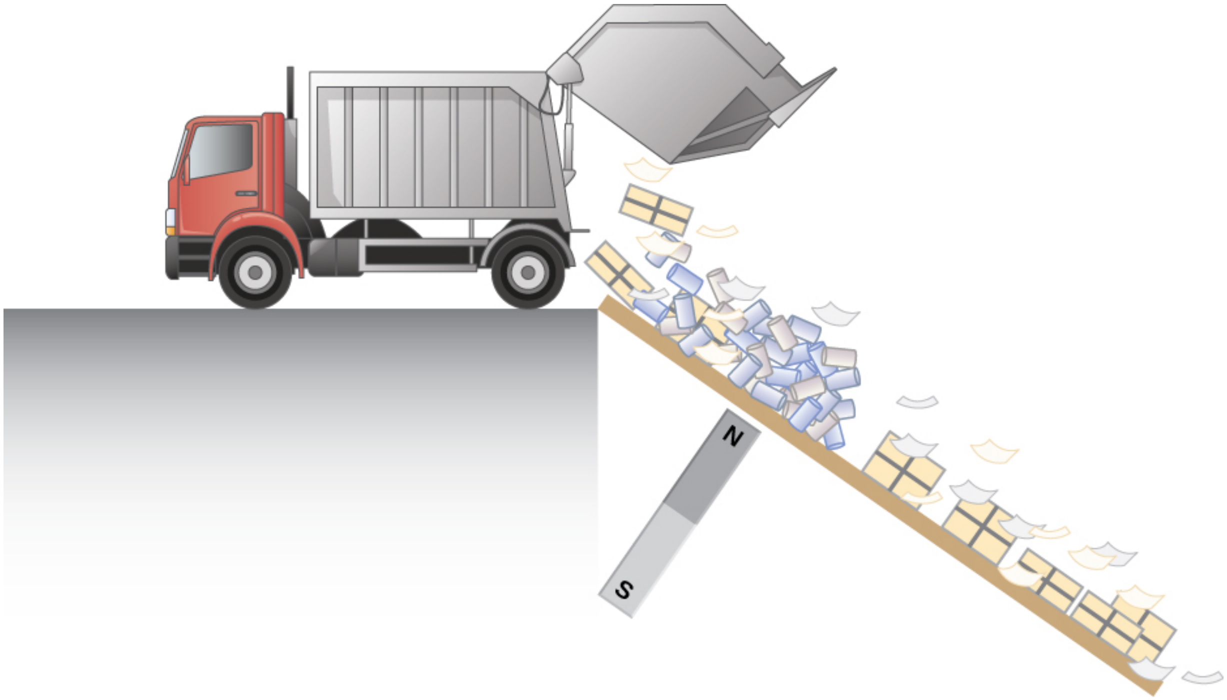
Applications of Magnetic Damping

One use of magnetic damping is found in sensitive laboratory balances. To have maximum sensitivity and accuracy, the balance must be as friction-free as possible. But if it is friction-free, then it will oscillate for a very long time. Magnetic damping is a simple and ideal solution. With magnetic damping, drag is proportional to speed and becomes zero at zero velocity. Thus, the oscillations are quickly damped, after which the damping force disappears, allowing the balance to be very sensitive ([link]). In most balances, magnetic damping is accomplished with a conducting disc that rotates in a fixed field.



Magnetic damping of this sensitive balance slows its oscillations. Since Faraday’s law of induction gives the greatest effect for the most rapid change, damping is greatest for large oscillations and goes to zero as the motion stops.

Since eddy currents and magnetic damping occur only in conductors, recycling centers can use magnets to separate metals from other materials. Trash is dumped in batches down a ramp, beneath which lies a powerful magnet. Conductors in the trash are slowed by magnetic damping while nonmetals in the trash move on, separating from the metals ([link]). This works for all metals, not just ferromagnetic ones. A magnet can separate out the ferromagnetic materials alone by acting on stationary trash.



Metals can be separated from other trash by magnetic drag. Eddy currents and magnetic drag are created in the metals sent down this ramp by the powerful magnet beneath it. Nonmetals move on.

Other major applications of eddy currents appear in metal detectors and braking systems in trains and roller coasters. Portable metal detectors ([link]) consist of a primary coil carrying an alternating current and a secondary coil in which a current is induced. An eddy current is induced in a piece of metal close to the detector, causing a change in the induced current within the secondary coil. This can trigger some sort of signal, such as a shrill noise.

A soldier in Iraq uses a metal detector to search for explosives and weapons. (credit: U.S. Army)

Braking using eddy currents is safer because factors such as rain do not affect the braking and the braking is smoother. However, eddy currents cannot bring the motion to a complete stop, since the braking force produced decreases as speed is reduced. Thus, speed can be reduced from say 20 m/s to 5 m/s, but another form of braking is needed to completely stop the vehicle. Generally, powerful rare-earth magnets such as neodymium magnets are used in roller coasters. [link] shows rows of magnets in such an application. The vehicle has metal fins (normally containing copper) that pass through the magnetic field, slowing the vehicle down in much the same way as with the pendulum bob shown in [link].

The rows of rare-earth magnets (protruding horizontally) are used for magnetic braking in roller coasters. (credit: Stefan Scheer)

Induction cooktops have electromagnets under their surface. The magnetic field is varied rapidly, producing eddy currents in the base of the pot, causing the pot and its contents to increase in temperature. Induction cooktops have high efficiencies and good response times but the base of the pot needs to be conductors, such as iron or steel, for induction to work.

Summary

- Current loops induced in moving conductors are called eddy currents. They can create significant drag, called magnetic damping.

- Manipulation of eddy currents has resulted in applications such as metal detectors, braking in trains or roller coasters, and induction cooktops.

Conceptual Questions

Exercise:

Problem:

A conducting sheet lies in a plane perpendicular to a magnetic field that is below the sheet. If oscillates at a high frequency and the conductor is made of a material of low resistivity, the region above the sheet is effectively shielded from . Explain why. Will the conductor shield this region from static magnetic fields?

Solution:

The conducting sheet is shielded from the changing magnetic fields by creating an induced emf. This induced emf creates an induced magnetic field that opposes any changes in magnetic fields from the field underneath. Therefore, there is no net magnetic field in the region above this sheet. If the field were due to a static magnetic field, no induced emf will be created since you need a changing magnetic flux to induce an emf. Therefore, this static magnetic field will not be shielded.

Exercise:

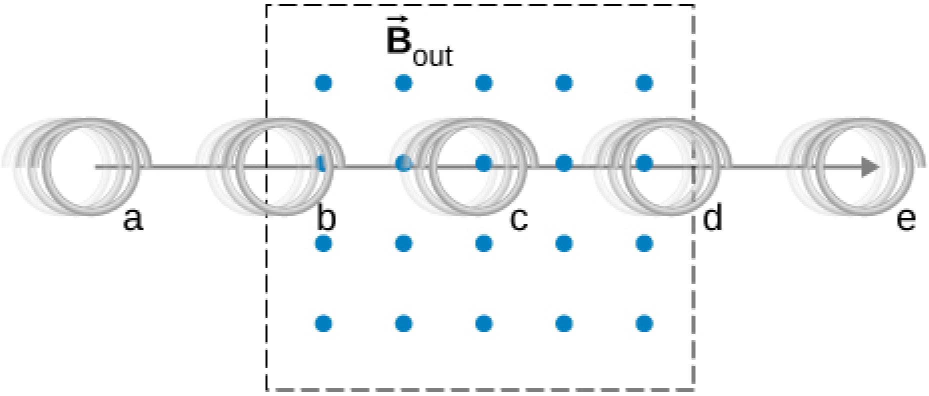
Problem:

Electromagnetic braking can be achieved by applying a strong magnetic field to a spinning metal disk attached to a shaft. (a) How can a magnetic field slow the spinning of a disk? (b) Would the brakes work if the disk was made of plastic instead of metal?

Exercise:

Problem:

A coil is moved through a magnetic field as shown below. The field is uniform inside the rectangle and zero outside. What is the direction of the induced current and what is the direction of the magnetic force on the coil at each position shown?



Solution:

a. zero induced current, zero force; b. clockwise induced current, force is to the left; c. zero induced current, zero force; d. counterclockwise induced current, force is to the left; e. zero induced current, zero force.

Glossary

magnetic damping
drag produced by eddy currents

eddy current
current loop in a conductor caused by motional emf

Electric Generators and Back Emf

By the end of this section, you will be able to:

- Explain how an electric generator works
- Determine the induced emf in a loop at any time interval, rotating at a constant rate in a magnetic field
- Show that rotating coils have an induced emf; in motors this is called back emf because it opposes the emf input to the motor

A variety of important phenomena and devices can be understood with Faraday’s law. In this section, we examine two of these.

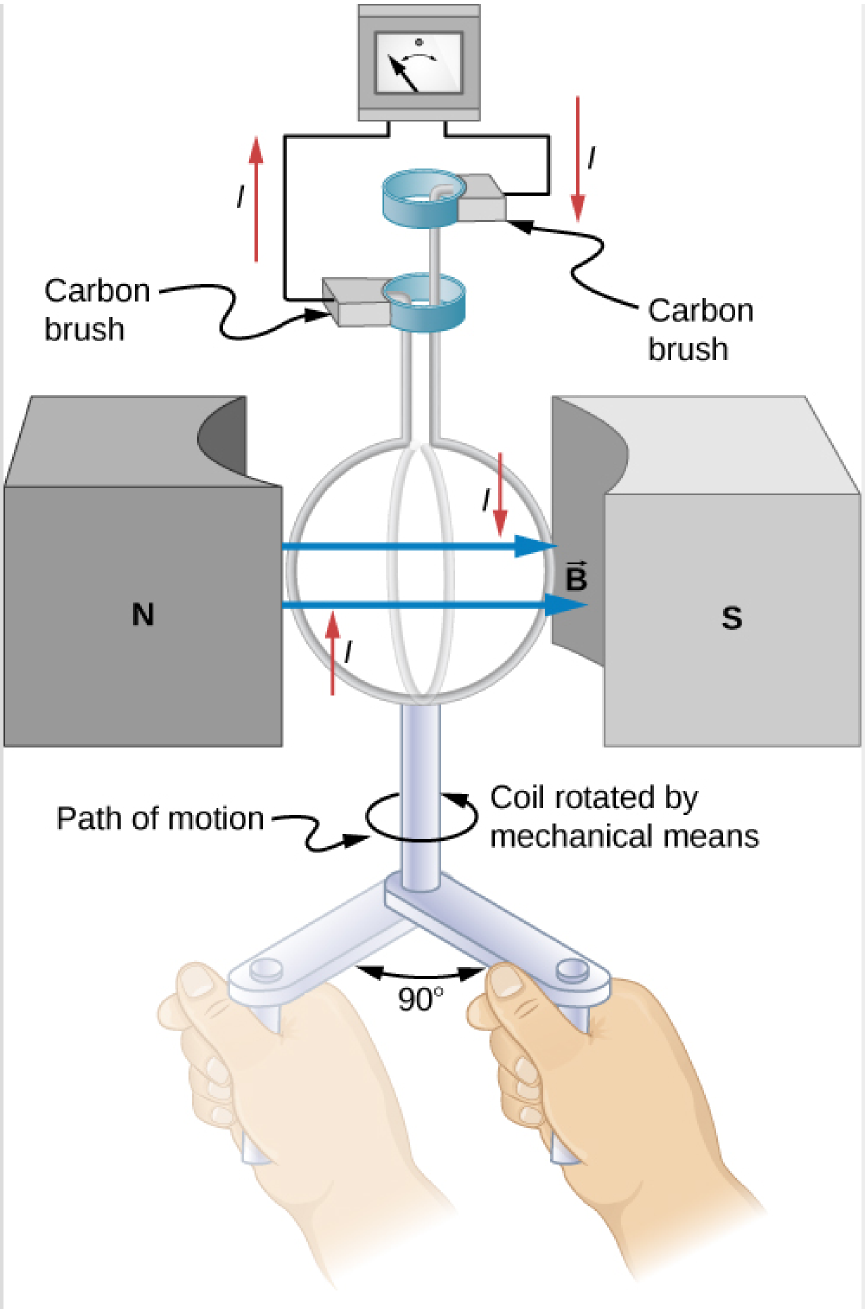
Electric Generators

Electric generators induce an emf by rotating a coil in a magnetic field, as briefly discussed in Motional Emf. We now explore generators in more detail. Consider the following example.

Example:

Calculating the Emf Induced in a Generator Coil

The generator coil shown in [link] is rotated through one-fourth of a revolution (from $\theta = 0^\circ$ to$\theta = 90^\circ$) in 15.0 ms. The 200-turn circular coil has a 5.00-cm radius and is in a uniform 0.80-T magnetic field. What is the emf induced?


Carbon brush
Carbon brush
I
I
I
I
N
S
$\vec{B}$
Path of motion
Coil rotated by mechanical means
90°

When this generator coil is rotated through one-fourth of a revolution, the magnetic flux Φ_m changes from its maximum to zero, inducing an emf.

Strategy

Faraday's law of induction is used to find the emf induced:

Equation:

$$\varepsilon = -N\frac{d\Phi_{\mathrm{m}}}{dt}.$$

We recognize this situation as the same one in [link]. According to the diagram, the projection of the surface normal vector $\hat{\mathbf{n}}$ to the magnetic field is initiallycos θ, and this is inserted by the definition of the dot product. The magnitude of the magnetic field and area of the loop are fixed over time, which makes the integration simplify quickly. The induced emf is written out using Faraday’s law:

Equation:

$$\varepsilon = NBA\sin\theta\frac{d\theta}{dt}.$$

Solution

We are given that $N = 200$, $B = 0.80\,\mathrm{T}$, $\theta = 90^\circ$, $d\theta = 90^\circ = \pi/2$, and $dt = 15.0\,\mathrm{ms}$. The area of the loop is

Equation:

$$A = \pi r^2 = (3.14)(0.0500\,\mathrm{m})^2 = 7.85 \times 10^{-3}\,\mathrm{m}^2.$$

Entering this value gives

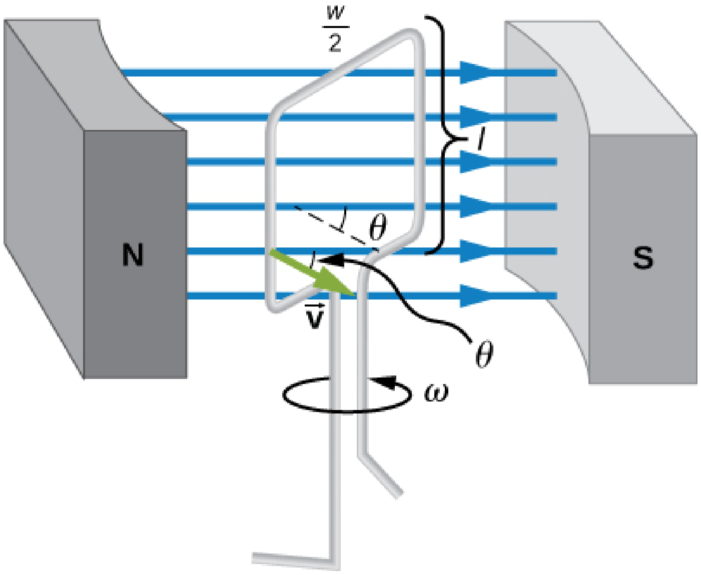
Equation:

$$\varepsilon = (200)(0.80\,\mathrm{T})(7.85 \times 10^{-3}\,\mathrm{m}^2)\sin(90^\circ)\,\frac{\pi/2}{15.0 \times 10^{-3}\,\mathrm{s}} = 131\,\mathrm{V}.$$

Significance

This is a practical average value, similar to the 120 V used in household power.

The emf calculated in [link] is the average over one-fourth of a revolution. What is the emf at any given instant? It varies with the angle between the magnetic field and a perpendicular to the coil. We can get an expression for emf as a function of time by considering the motional emf on a rotating rectangular coil of width w and height l in a uniform magnetic field, as illustrated in [link].



A generator with a single rectangular coil rotated at constant angular velocity in a uniform magnetic field produces an emf that varies sinusoidally in time. Note the generator is similar to a motor, except the shaft is rotated to produce a current rather than the other way around.

Charges in the wires of the loop experience the magnetic force, because they are moving in a magnetic field. Charges in the vertical wires experience forces parallel to the wire, causing currents. But those in the top and bottom segments feel a force perpendicular to the wire, which does not cause a current. We can thus find the induced emf by considering only the side wires. Motional emf is given to be $\varepsilon = Blv$, where the velocity v is perpendicular to the magnetic field B. Here the velocity is at an angle θ with B, so that its component perpendicular to B is $v \sin \theta$ (see [link]). Thus, in this case, the emf induced on each side is $\varepsilon = Blv \sin \theta$, and they are in the same direction. The total emf around the loop is then

Equation:

$$\varepsilon = 2Blv \sin \theta.$$

This expression is valid, but it does not give emf as a function of time. To find the time dependence of emf, we assume the coil rotates at a constant angular velocity ω. The angle θ is related to angular velocity by $\theta = \omega t$, so that

Equation:

$$\varepsilon = 2Blv\sin(\omega t).$$

Now, linear velocity *v* is related to angular velocity ω by $v = r\omega$. Here, $r = w/2$, so that $v = (w/2)\omega$, and

Equation:

$$\varepsilon = 2Bl\frac{w}{2}\omega\sin\omega t = (lw)B\omega\sin\omega t.$$

Noting that the area of the loop is $A = lw$, and allowing for *N* loops, we find that

Note:

Equation:

$$\varepsilon = NBA\omega\sin(\omega t).$$

This is the emf induced in a generator coil of *N* turns and area *A* rotating at a constant angular velocity ω in a uniform magnetic field *B*. This can also be expressed as

Equation:

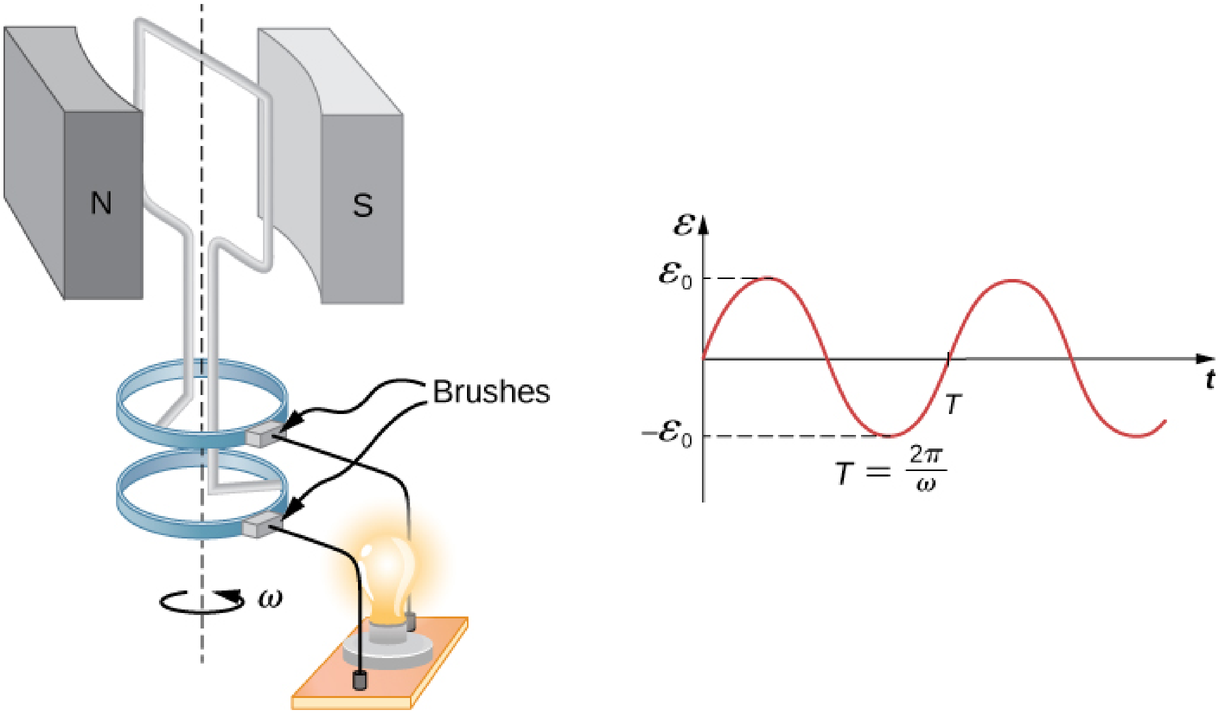
$$\varepsilon = \varepsilon_0\sin\omega t,$$

where

Equation:

$$\varepsilon_0 = NAB\omega$$

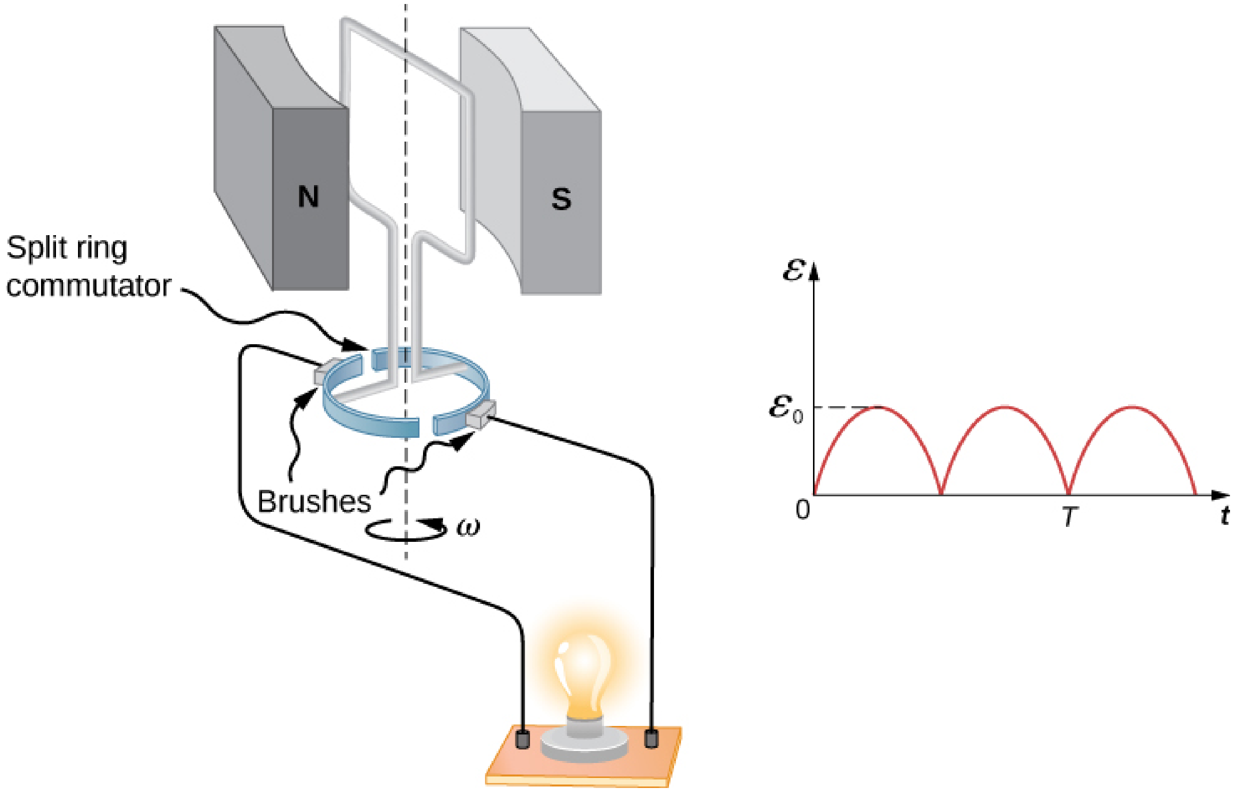
is the peak emf, since the maximum value of $\sin(wt) = 1$. Note that the frequency of the oscillation is $f = \omega/2\pi$ and the period is $T = 1/f = 2\pi/\ \omega$. [link] shows a graph of emf as a function of time, and it now seems reasonable that ac voltage is sinusoidal.



The emf of a generator is sent to a light bulb with the system of rings and brushes shown. The graph gives the emf of the generator as a function of time, where ε_0 is the peak emf. The period is $T = 1/f = 2\pi/\omega$, where f is the frequency.

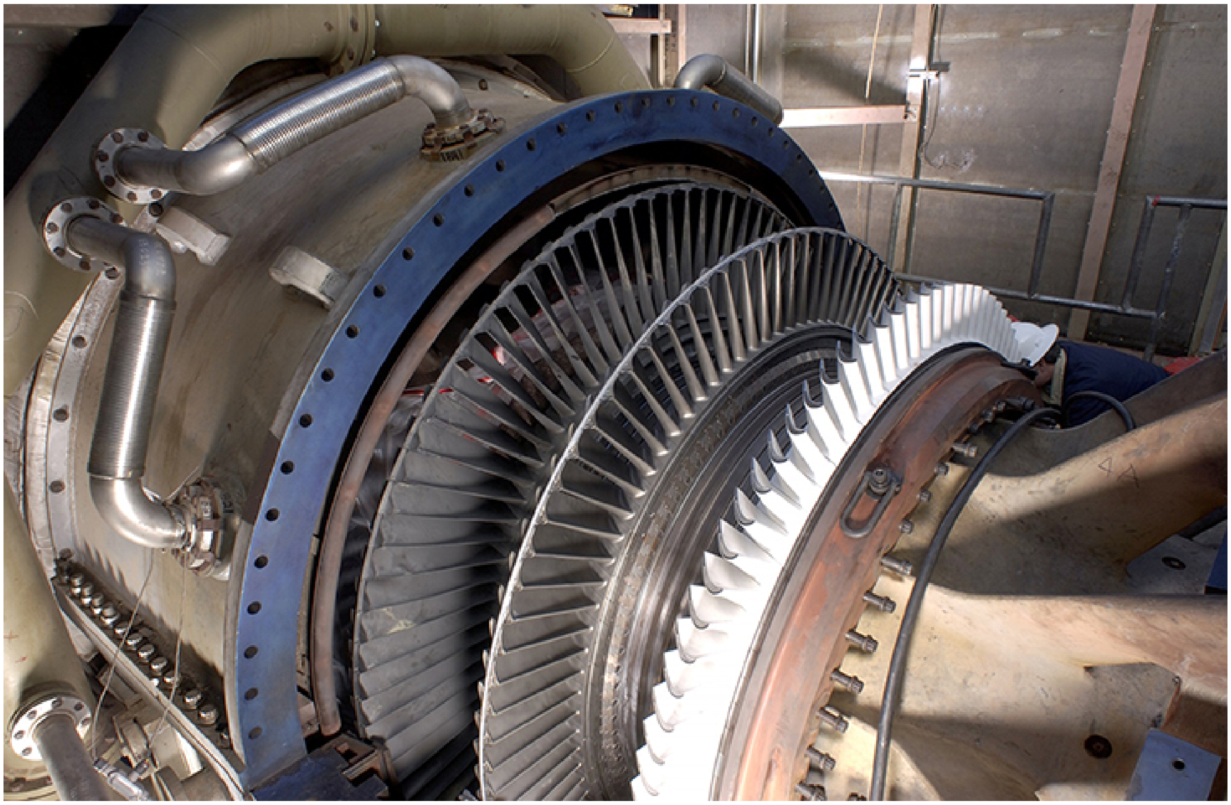
The fact that the peak emf is $\varepsilon_0 = NBA\omega$ makes good sense. The greater the number of coils, the larger their area, and the stronger the field, the greater the output voltage. It is interesting that the faster the generator is spun (greater ω), the greater the emf. This is noticeable on bicycle generators—at least the cheaper varieties.

[link] shows a scheme by which a generator can be made to produce pulsed dc. More elaborate arrangements of multiple coils and split rings can produce smoother dc, although electronic rather than mechanical means are usually used to make ripple-free dc.



Split rings, called commutators, produce a pulsed dc emf output in this configuration.

In real life, electric generators look a lot different from the figures in this section, but the principles are the same. The source of mechanical energy that turns the coil can be falling water (hydropower), steam produced by the burning of fossil fuels, or the kinetic energy of wind. [link] shows a cutaway view of a steam turbine; steam moves over the blades connected to the shaft, which rotates the coil within the generator. The generation of electrical energy from mechanical energy is the basic principle of all power that is sent through our electrical grids to our homes.



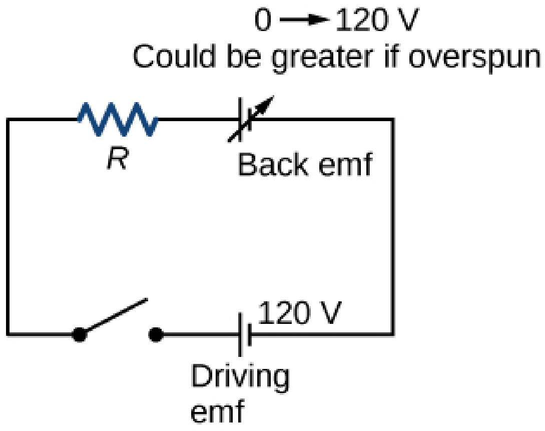
Steam turbine/generator. The steam produced by burning coal impacts the turbine blades, turning the shaft, which is connected to the generator.

Generators illustrated in this section look very much like the motors illustrated previously. This is not coincidental. In fact, a motor becomes a generator when its shaft rotates. Certain early automobiles used their starter motor as a generator. In the next section, we further explore the action of a motor as a generator.

Back Emf

Generators convert mechanical energy into electrical energy, whereas motors convert electrical energy into mechanical energy. Thus, it is not surprising that motors and generators have the same general construction. A motor works by sending a current through a loop of wire located in a magnetic field. As a result, the magnetic field exerts torque on the loop. This rotates a shaft, thereby extracting mechanical work out of the electrical current sent in initially. (Refer to Force and Torque on a Current Loop for a discussion on motors that will help you understand more about them before proceeding.)

When the coil of a motor is turned, magnetic flux changes through the coil, and an emf (consistent with Faraday’s law) is induced. The motor thus acts as a generator whenever its coil rotates. This happens whether the shaft is turned by an external input, like a belt drive, or by the action of the motor itself. That is, when a motor is doing work and its shaft is turning, an emf is generated. Lenz’s law tells us the emf opposes any change, so that the input emf that powers the motor is opposed by the motor’s self-generated emf, called the **back emf** of the motor ([link]).



The coil of a dc motor is represented as a resistor in this schematic. The back emf is represented as a variable emf that opposes the emf driving the motor. Back emf is zero when the motor is not turning and increases proportionally to the motor’s angular velocity.

The generator output of a motor is the difference between the supply voltage and the back emf. The back emf is zero when the motor is first turned on, meaning that the coil receives the full driving voltage and the motor draws maximum current when it is on but not turning. As the motor turns faster, the back emf grows, always opposing the driving emf, and reduces both the voltage across the coil and the amount of current it draws. This effect is noticeable in many common situations. When a vacuum cleaner, refrigerator, or washing machine is first turned on, lights in the same circuit dim briefly due to the *IR* drop produced in feeder lines by the large current drawn by the motor.

When a motor first comes on, it draws more current than when it runs at its normal operating speed. When a mechanical load is placed on the motor, like an electric wheelchair going up a hill, the motor slows, the back emf drops, more current flows, and more work can be done. If the motor runs at too low a speed, the larger current can overheat it (via resistive power in the coil, $P=I^2R$), perhaps even burning it out. On the other hand, if there is no mechanical load on the motor, it increases its angular velocity ω until the back emf is nearly equal to the driving emf. Then the motor uses only enough energy to overcome friction.

Eddy currents in iron cores of motors can cause troublesome energy losses. These are usually minimized by constructing the cores out of thin, electrically insulated sheets of iron. The magnetic properties of the core are hardly affected by the lamination of the insulating sheet, while the resistive heating is reduced considerably. Consider, for example, the motor coils represented in [link]. The coils have an equivalent resistance of $0.400\ \Omega$ and are driven by an emf of 48.0 V. Shortly after being turned on, they draw a current

Equation:

$$I=V/R=(48.0\text{ V})/(0.400\ \Omega)=120\text{ A}$$

and thus dissipate $P=I^2R=5.76\text{ kW}$ of energy as heat transfer. Under normal operating conditions for this motor, suppose the back emf is 40.0 V. Then at operating speed, the total voltage across the coils is 8.0 V (48.0 V minus the 40.0 V back emf), and the current drawn is

Equation:

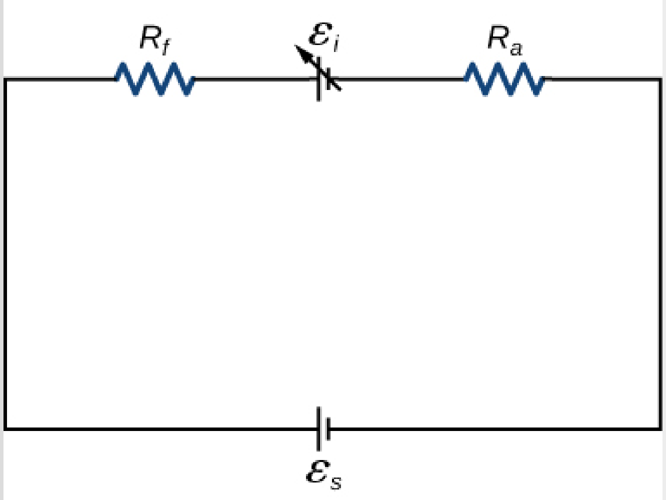
$$I=V/R=(8.0\text{ V})/(0.400\ \Omega)=20\text{ A}.$$

Under normal load, then, the power dissipated is $P=IV=(20\text{ A})(8.0\text{ V})=160\text{ W}$. This does not cause a problem for this motor, whereas the former 5.76 kW would burn out the coils if sustained.

Example:

A Series-Wound Motor in Operation

The total resistance (R_f+R_a) of a series-wound dc motor is $2.0\ \Omega$ ([link]). When connected to a 120-V source (ε_S), the motor draws 10 A while running at constant angular velocity. (a) What is the back emf induced in the rotating coil, ε_i? (b) What is the mechanical power output of the motor? (c) How much power is dissipated in the resistance of the coils? (d) What is the power output of the 120-V source? (e) Suppose the load on the motor increases, causing it to slow down to the point where it draws 20 A. Answer parts (a) through (d) for this situation.



Circuit representation of a series-wound direct current motor.

Strategy
The back emf is calculated based on the difference between the supplied voltage and the loss from the current through the resistance. The power from each device is calculated from one of the power formulas based on the given information.
Solution

a. The back emf is
   Equation:

$$\varepsilon_i = \varepsilon_s - I(R_f + R_a) = 120\text{ V} - (10\text{ A})(2.0\,\Omega) = 100\text{ V}.$$

b. Since the potential across the armature is 100 V when the current through it is 10 A, the power output of the motor is
   Equation:

$$P_{\text{m}} = \varepsilon_i I = (100\text{ V})(10\text{ A}) = 1.0 \times 10^3\text{ W}.$$

c. A 10-A current flows through coils whose combined resistance is $2.0\,\Omega$, so the power dissipated in the coils is
   Equation:

$$P_R = I^2R = (10\text{ A})^2(2.0\,\Omega) = 2.0 \times 10^2\text{ W}.$$

d. Since 10 A is drawn from the 120-V source, its power output is

Equation:

$$P_s = \varepsilon_s I = (120\text{ V})(10\text{ A}) = 1.2\ \times\ 10^3\text{ W}.$$

e. Repeating the same calculations with $I = 20\text{ A}$, we find
Equation:

$$\varepsilon_{\text{i}} =\ 80\text{ V}, P_{\text{m}} =\ 1.6\ \times\ 10^3\text{ W}, P_R = 8.0\ \times\ 10^2\text{ W},\ \text{and}\ P_s = 2.4\ \times\ 10^3\text{ W}.$$

The motor is turning more slowly in this case, so its power output and the power of the source are larger.

Significance
Notice that we have an energy balance in part (d):
$1.2\ \times\ 10^3\text{ W} = 1.0\ \times\ 10^3\text{ W} + 2.0\ \times\ 10^2\text{ W}.$

Summary

- An electric generator rotates a coil in a magnetic field, inducing an emf given as a function of time by $\varepsilon = NBA\omega\,\text{sin}(\omega t)$ where *A* is the area of an *N*-turn coil rotated at a constant angular velocity ω in a uniform magnetic field $\mathbf{B}$.
- The peak emf of a generator is $\varepsilon_0 = NBA\omega$.
- Any rotating coil produces an induced emf. In motors, this is called back emf because it opposes the emf input to the motor.

Problems

Exercise:

Problem:

Design a current loop that, when rotated in a uniform magnetic field of strength 0.10 T, will produce an emf $\varepsilon = \varepsilon_0\sin\omega t$, where $\varepsilon_0 = 110\text{ V}$ and $\omega = 120\pi\,\text{rad/s}$.

Solution:

three turns with an area of 1 m^2

Exercise:

Problem:

A flat, square coil of 20 turns that has sides of length 15.0 cm is rotating in a magnetic field of strength 0.050 T. If the maximum emf produced in the coil is 30.0 mV, what is the angular velocity of the coil?

Exercise:

Problem:

A 50-turn rectangular coil with dimensions $0.15\,\text{m} \times 0.40\,\text{m}$ rotates in a uniform magnetic field of magnitude 0.75 T at 3600 rev/min. (a) Determine the emf induced in the coil as a function of time. (b) If the coil is connected to a 1000-Ω resistor, what is the power as a function of time required to keep the coil turning at 3600 rpm? (c) Answer part (b) if the coil is connected to a 2000-Ω resistor.

Solution:

a. $$\begin{aligned} \omega &= 120\pi\,\text{rad/s}, \\ \varepsilon &= 850\sin 120\,\pi\text{t V}; \end{aligned}$$
b. $P = 720\sin^2 120\,\pi t$ W;
c. $P = 360\sin^2 120\,\pi t$ W

Exercise:

Problem:

The square armature coil of an alternating current generator has 200 turns and is 20.0 cm on side. When it rotates at 3600 rpm, its peak output voltage is 120 V. (a) What is the frequency of the output voltage? (b) What is the strength of the magnetic field in which the coil is turning?

Exercise:

Problem:

A flip coil is a relatively simple device used to measure a magnetic field. It consists of a circular coil of *N* turns wound with fine conducting wire. The coil is attached to a ballistic galvanometer, a device that measures the total charge that passes through it. The coil is placed in a magnetic field $\mathbf{B}$ such that its face is perpendicular to the field. It is then flipped through 180°, and the total charge *Q* that flows through the galvanometer is measured. (a) If the total resistance of the coil and galvanometer is *R*, what is the relationship between *B* and *Q*? Because the coil is very small, you can assume that $\mathbf{B}$ is uniform over it. (b) How can you determine whether or not the magnetic field is perpendicular to the face of the coil?

Solution:

a. *B* is proportional to *Q*; b. If the coin turns easily, the magnetic field is perpendicular. If the coin is at an equilibrium position, it is parallel.

Exercise:

Problem:

The flip coil of the preceding problem has a radius of 3.0 cm and is wound with 40 turns of copper wire. The total resistance of the coil and ballistic galvanometer is $0.20\,\Omega$. When the coil is flipped through 180° in a magnetic field $\mathbf{B}$, a change of 0.090 C flows through the ballistic galvanometer. (a) Assuming that $\mathbf{B}$ and the face of the coil are initially perpendicular, what is the magnetic field? (b) If the coil is flipped through 90°, what is the reading of the galvanometer?

Exercise:

Problem:

A 120-V, series-wound motor has a field resistance of 80 Ω and an armature resistance of 10 Ω. When it is operating at full speed, a back emf of 75 V is generated. (a) What is the initial current drawn by the motor? When the motor is operating at full speed, where are (b) the current drawn by the motor, (c) the power output of the source, (d) the power output of the motor, and (e) the power dissipated in the two resistances?

Solution:

a. 1.33 A; b. 0.50 A; c. 60 W; d. 37.5 W; e. 22.5W

Exercise:

Problem:

A small series-wound dc motor is operated from a 12-V car battery. Under a normal load, the motor draws 4.0 A, and when the armature is clamped so that it cannot turn, the motor draws 24 A. What is the back emf when the motor is operating normally?

Glossary

back emf
: emf generated by a running motor, because it consists of a coil turning in a magnetic field; it opposes the voltage powering the motor

peak emf	maximum emf produced by a generator
electric generator	device for converting mechanical work into electric energy; it induces an emf by rotating a coil in a magnetic field

Applications of Electromagnetic Induction

By the end of this section, you will be able to:

- Explain how computer hard drives and graphic tablets operate using magnetic induction
- Explain how hybrid/electric vehicles and transcranial magnetic stimulation use magnetic induction to their advantage

Modern society has numerous applications of Faraday’s law of induction, as we will explore in this chapter and others. At this juncture, let us mention several that involve recording information using magnetic fields.

Some computer hard drives apply the principle of magnetic induction. Recorded data are made on a coated, spinning disk. Historically, reading these data was made to work on the principle of induction. However, most input information today is carried in digital rather than analog form—a series of 0s or 1s are written upon the spinning hard drive. Therefore, most hard drive readout devices do not work on the principle of induction, but use a technique known as giant magnetoresistance. Giant magnetoresistance is the effect of a large change of electrical resistance induced by an applied magnetic field to thin films of alternating ferromagnetic and nonmagnetic layers. This is one of the first large successes of nanotechnology.

Graphics tablets, or tablet computers where a specially designed pen is used to draw digital images, also applies induction principles. The tablets discussed here are labeled as passive tablets, since there are other designs that use either a battery-operated pen or optical signals to write with. The passive tablets are different than the touch tablets and phones many of us use regularly, but may still be found when signing your signature at a cash register. Underneath the screen, shown in [link], are tiny wires running across the length and width of the screen. The pen has a tiny magnetic field coming from the tip. As the tip brushes across the screen, a changing magnetic field is felt in the wires which translates into an induced emf that is converted into the line you just drew.

A tablet with a specially designed pen to write with is another application of magnetic induction. (credit: Jane Whitney)

Another application of induction is the magnetic stripe on the back of your personal credit card as used at the grocery store or the ATM machine. This works on the same principle as the audio or video tape, in which a playback head reads personal information from your card.

Note:
Check out this video to see how flashlights can use magnetic induction. A magnet moves by your mechanical work through a wire. The induced

current charges a capacitor that stores the charge that will light the lightbulb even while you are not doing this mechanical work.

Electric and hybrid vehicles also take advantage of electromagnetic induction. One limiting factor that inhibits widespread acceptance of 100% electric vehicles is that the lifetime of the battery is not as long as the time you get to drive on a full tank of gas. To increase the amount of charge in the battery during driving, the motor can act as a generator whenever the car is braking, taking advantage of the back emf produced. This extra emf can be newly acquired stored energy in the car’s battery, prolonging the life of the battery.

Another contemporary area of research in which electromagnetic induction is being successfully implemented is transcranial magnetic stimulation (TMS). A host of disorders, including depression and hallucinations, can be traced to irregular localized electrical activity in the brain. In transcranial magnetic stimulation, a rapidly varying and very localized magnetic field is placed close to certain sites identified in the brain. The usage of TMS as a diagnostic technique is well established.

Note:
Check out this Youtube video to see how rock-and-roll instruments like electric guitars use electromagnetic induction to get those strong beats.

Summary

- Hard drives utilize magnetic induction to read/write information.
- Other applications of magnetic induction can be found in graphics tablets, electric and hybrid vehicles, and in transcranial magnetic stimulation.

Key Equations

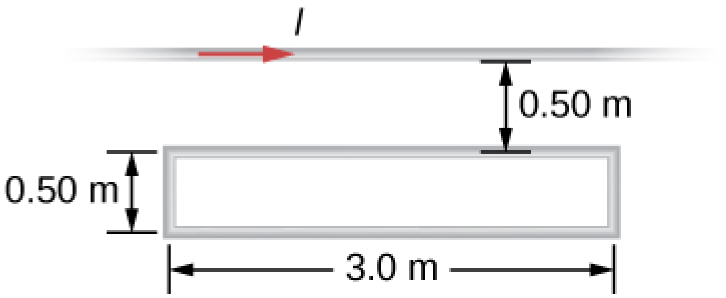
Magnetic flux	$\Phi_{\mathrm{m}} = \int_S \vec{\mathbf{B}} \cdot \hat{\mathbf{n}} dA$
Faraday’s law	$\varepsilon = -N \frac{d\Phi_{\mathrm{m}}}{dt}$
Motionally induced emf	$\varepsilon = Blv$
Motional emf around a circuit	$\varepsilon = \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = -\frac{d\Phi_{\mathrm{m}}}{dt}$
Emf produced by an electric generator	$\varepsilon = NBA\,\omega\ \sin(\omega t)$

Additional Problems

Exercise:

Problem:

Shown in the following figure is a long, straight wire and a single-turn rectangular loop, both of which lie in the plane of the page. The wire is parallel to the long sides of the loop and is 0.50 m away from the closer side. At an instant when the emf induced in the loop is 2.0 V, what is the time rate of change of the current in the wire?



Solution:

4.8×10^6 A/s

Exercise:

Problem:

A metal bar of mass 500 g slides outward at a constant speed of 1.5 cm/s over two parallel rails separated by a distance of 30 cm which are part of a U-shaped conductor. There is a uniform magnetic field of magnitude 2 T pointing out of the page over the entire area. The railings and metal bar have an equivalent resistance of $150\ \Omega$. (a) Determine the induced current, both magnitude and direction. (b) Find the direction of the induced current if the magnetic field is pointing into the page. (c) Find the direction of the induced current if the magnetic field is pointed into the page and the bar moves inwards.

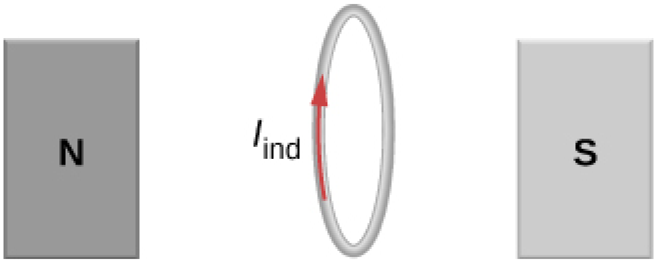
Exercise:

Problem:

A current is induced in a circular loop of radius 1.5 cm between two poles of a horseshoe electromagnet when the current in the electromagnet is varied. The magnetic field in the area of the loop is perpendicular to the area and has a uniform magnitude. If the rate of change of magnetic field is 10 T/s, find the magnitude and direction of the induced current if resistance of the loop is $25\ \Omega$.

Solution:

2.83×10^{-4} A, the direction as follows for increasing magnetic field:



Exercise:

Problem:

A metal bar of length 25 cm is placed perpendicular to a uniform magnetic field of strength 3 T. (a) Determine the induced emf between the ends of the rod when it is not moving. (b) Determine the emf when the rod is moving perpendicular to its length and magnetic field with a speed of 50 cm/s.

Exercise:

Problem:

A coil with 50 turns and area 10 cm^2 is oriented with its plane perpendicular to a 0.75-T magnetic field. If the coil is flipped over (rotated through 180°) in 0.20 s, what is the average emf induced in it?

Solution:

0.375 V

Exercise:

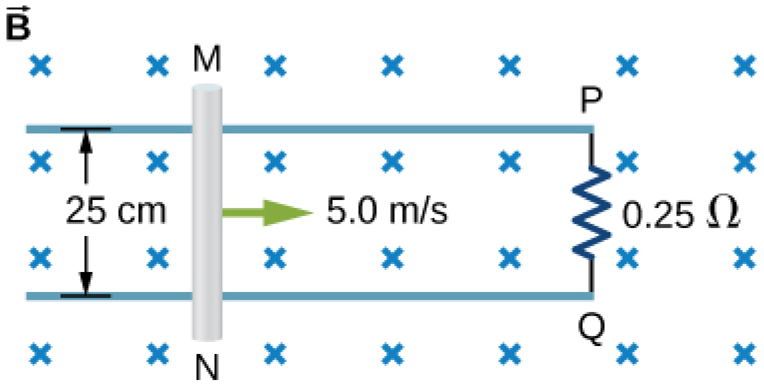
Problem:

A 2-turn planer loop of flexible wire is placed inside a long solenoid of n turns per meter that carries a constant current I_0. The area A of the loop is changed by pulling on its sides while ensuring that the plane of the loop always remains perpendicular to the axis of the solenoid. If $n = 500$ turns per meter, $I_0 = 20\text{ A}$, and $A = 20\text{ cm}^2$, what is the emf induced in the loop when $dA/dt = 100$?

Exercise:

Problem:

The conducting rod shown in the accompanying figure moves along parallel metal rails that are 25-cm apart. The system is in a uniform magnetic field of strength 0.75 T, which is directed into the page. The resistances of the rod and the rails are negligible, but the section PQ has a resistance of $0.25\ \Omega$. (a) What is the emf (including its sense) induced in the rod when it is moving to the right with a speed of 5.0 m/s? (b) What force is required to keep the rod moving at this speed? (c) What is the rate at which work is done by this force? (d) What is the power dissipated in the resistor?



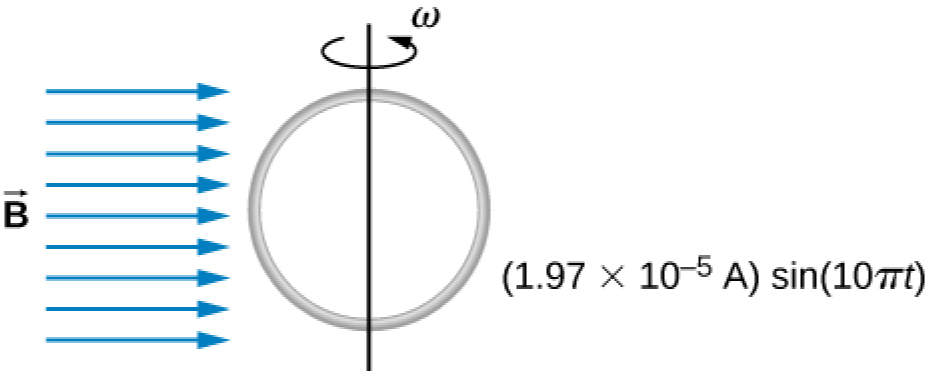
Solution:

a. 0.94 V; b. 0.70 N; c. 3.52 J/s; d. 3.52 W

Exercise:

Problem:

A circular loop of wire of radius 10 cm is mounted on a vertical shaft and rotated at a frequency of 5 cycles per second in a region of uniform magnetic field of 2 Gauss perpendicular to the axis of rotation. (a) Find an expression for the time-dependent flux through the ring. (b) Determine the time-dependent current through the ring if it has a resistance of 10 Ω.



Exercise:

Problem:

The magnetic field between the poles of a horseshoe electromagnet is uniform and has a cylindrical symmetry about an axis from the middle of the South Pole to the middle of the North Pole. The magnitude of the magnetic field changes as a rate of dB/dt due to the changing current through the electromagnet. Determine the electric field at a distance r from the center.

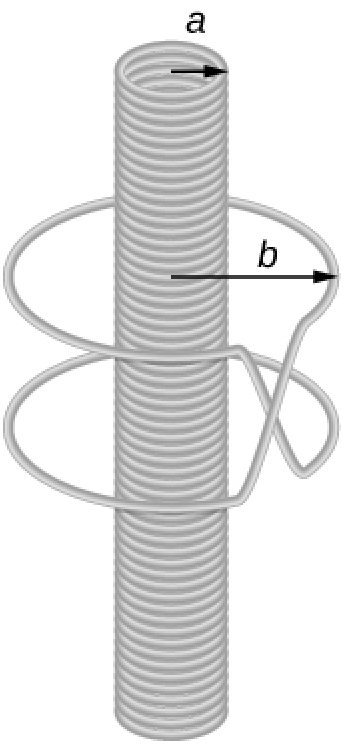
Solution:

$\left(\frac{dB}{dt}\right)\frac{A}{2\pi r}$

Exercise:

Problem:

A long solenoid of radius a with n turns per unit length is carrying a time-dependent current $I(t) = I_0 \sin(\omega t)$, where I_0 and ω are constants. The solenoid is surrounded by a wire of resistance R that has two circular loops of radius b with $b > a$ (see the following figure). Find the magnitude and direction of current induced in the outer loops at time $t = 0$.



Exercise:

Problem:

A 120-V, series-wound dc motor draws 0.50 A from its power source when operating at full speed, and it draws 2.0 A when it starts. The resistance of the armature coils is $10\,\Omega$. (a) What is the resistance of the field coils? (b) What is the back emf of the motor when it is running at full speed? (c) The motor operates at a different speed and draws 1.0 A from the source. What is the back emf in this case?

Solution:

a. $R_\text{f}+R_\text{a}=\frac{120\text{ V}}{2.0\text{ A}}=60\,\Omega,$ so $R_\text{f}=50\,\Omega;$
b. $I=\frac{\varepsilon_\text{s}-\varepsilon_\text{i}}{R_\text{f}+R_\text{a}},\Rightarrow\varepsilon_\text{i}=90\text{ V};$
c. $\varepsilon_\text{i}=60\text{ V}$

Exercise:

Problem:

The armature and field coils of a series-wound motor have a total resistance of $3.0\,\Omega$. When connected to a 120-V source and running at normal speed, the motor draws 4.0 A. (a) How large is the back emf? (b) What current will the motor draw just after it is turned on? Can you suggest a way to avoid this large initial current?

Challenge Problems

Exercise:

Problem:

A copper wire of length *L* is fashioned into a circular coil with *N* turns. When the magnetic field through the coil changes with time, for what value of *N* is the induced emf a maximum?

Solution:

N is a maximum number of turns allowed.

Exercise:

Problem:

A 0.50-kg copper sheet drops through a uniform horizontal magnetic field of 1.5 T, and it reaches a terminal velocity of 2.0 m/s. (a) What is the net magnetic force on the sheet after it reaches terminal velocity? (b) Describe the mechanism responsible for this force. (c) How much power is dissipated as Joule heating while the sheet moves at terminal velocity?

Exercise:

Problem:

A circular copper disk of radius 7.5 cm rotates at 2400 rpm around the axis through its center and perpendicular to its face. The disk is in a uniform magnetic field $\vec{\mathbf{B}}$ of strength 1.2 T that is directed along the axis. What is the potential difference between the rim and the axis of the disk?

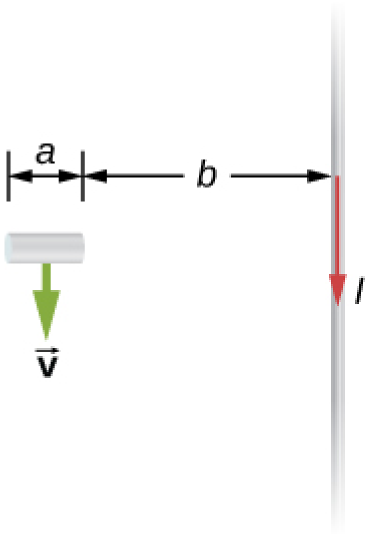
Solution:

0.848 V

Exercise:

Problem:

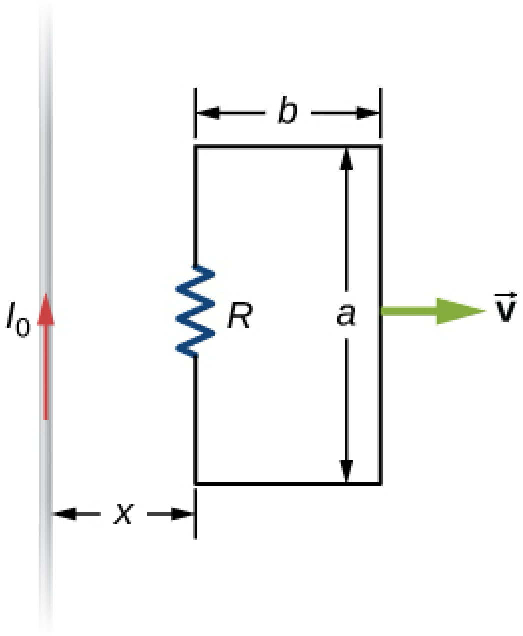
A short rod of length *a* moves with its velocity $\vec{\mathbf{v}}$ parallel to an infinite wire carrying a current *I* (see below). If the end of the rod nearer the wire is a distance *b* from the wire, what is the emf induced in the rod?



Exercise:

Problem:

A rectangular circuit containing a resistance *R* is pulled at a constant velocity $\vec{\mathbf{v}}$ away from a long, straight wire carrying a current I_0 (see below). Derive an equation that gives the current induced in the circuit as a function of the distance *x* between the near side of the circuit and the wire.



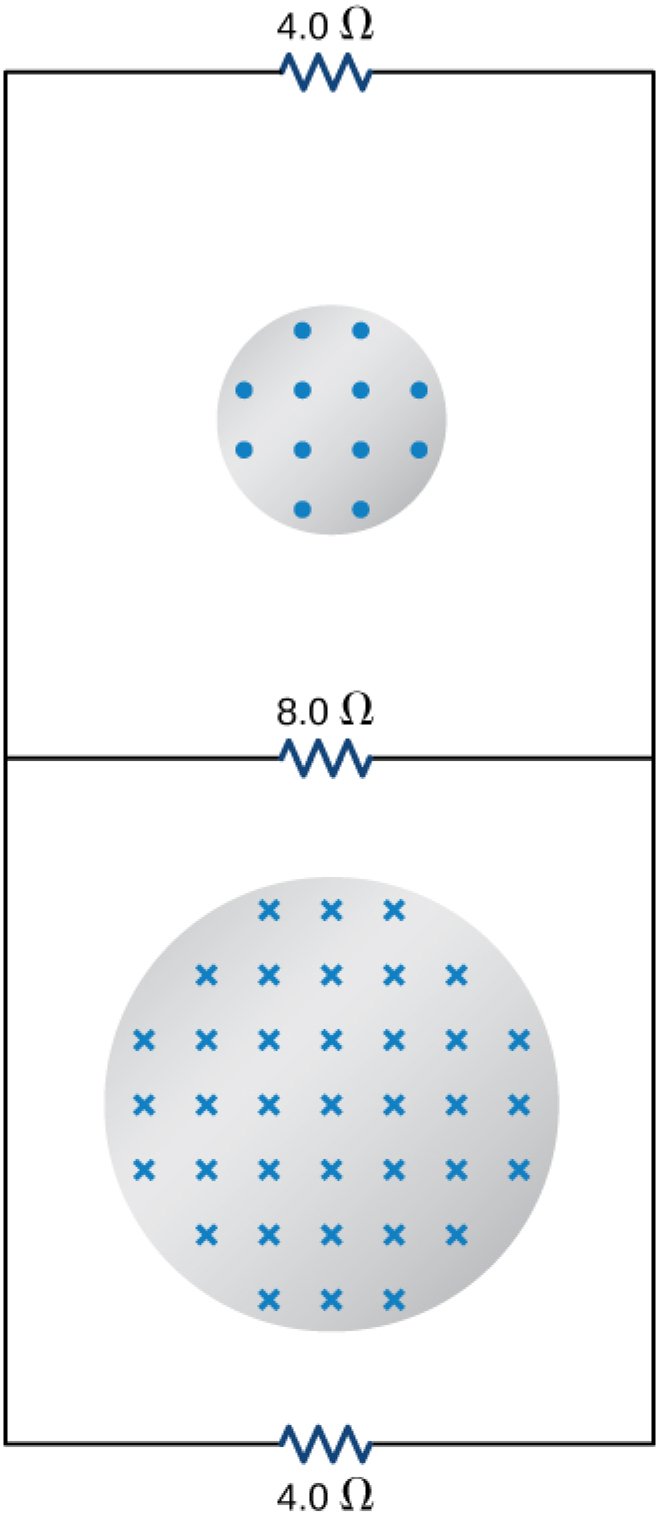
Solution:

$$\begin{aligned} \Phi &= \frac{\mu_0 I_0 a}{2\pi} \ln\left(1+\frac{b}{x}\right), \quad \varepsilon = \frac{\mu_0 I_0 abv}{2\pi x(x+b)}, \\ \text{so } I &= \frac{\mu_0 I_0 abv}{2\pi Rx(x+b)} \end{aligned}$$

Exercise:

Problem:

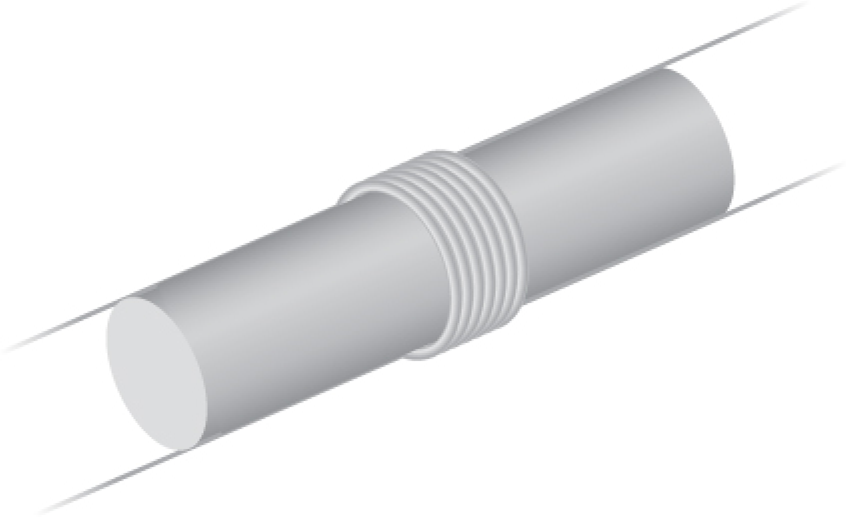
Two infinite solenoids cross the plane of the circuit as shown below. The radii of the solenoids are 0.10 and 0.20 m, respectively, and the current in each solenoid is changing such that $dB/dt = 50.0\ \mathrm{T/s}$. What are the currents in the resistors of the circuit?



Exercise:

Problem:

An eight-turn coil is *tightly wrapped* around the outside of the long solenoid as shown below. The radius of the solenoid is 2.0 cm and it has 10 turns per centimeter. The current through the solenoid increases according to $I = I_0(1 - e^{-\alpha t})$, where $I_0 = 4.0\text{ A}$ and $\alpha = 2.0\ \times\ 10^{-2}\text{ s}^{-1}$. What is the emf induced in the coil when (a) $t = 0$, (b) $t = 1.0\ \times\ 10^2\text{ s}$, and (c) $t \rightarrow \infty$?



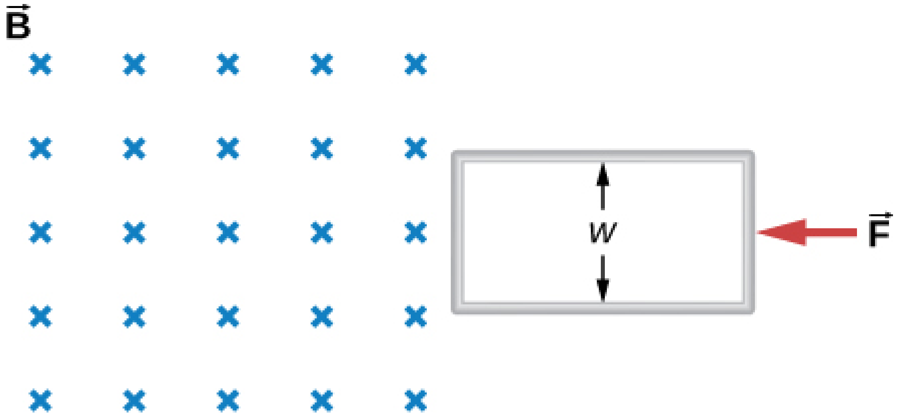
Solution:

a. $1.01\ \times\ 10^{-6}$V; b. $1.37\ \times\ 10^{-7}$ V; c. 0 V

Exercise:

Problem:

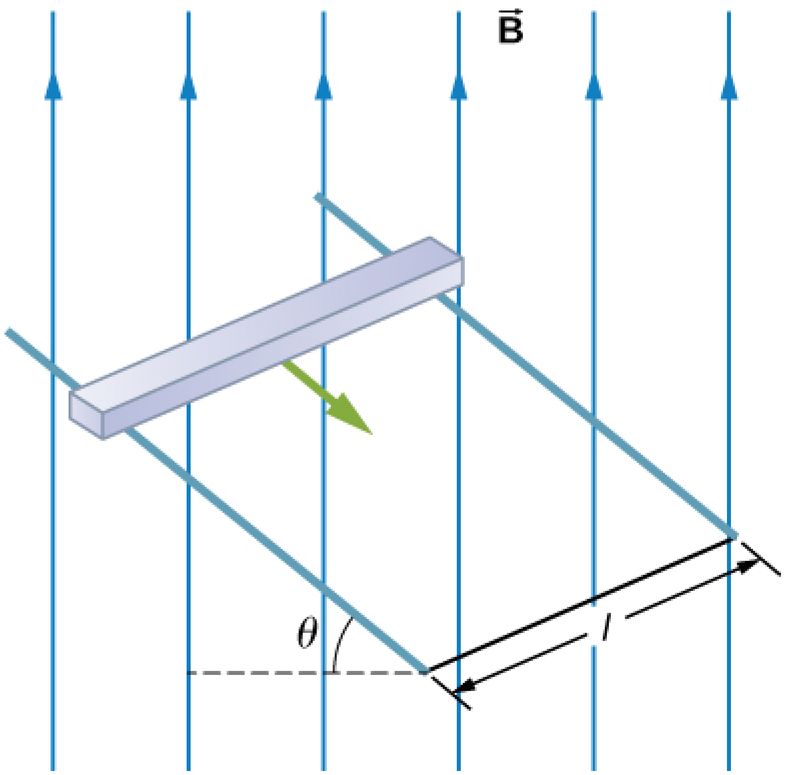
Shown below is a long rectangular loop of width *w*, length *l*, mass *m*, and resistance *R*. The loop starts from rest at the edge of a uniform magnetic field $\vec{\mathbf{B}}$ and is pushed into the field by a constant force $\vec{\mathbf{F}}$. Calculate the speed of the loop as a function of time.



Exercise:

Problem:

A square bar of mass m and resistance R is sliding without friction down very long, parallel conducting rails of negligible resistance (see below). The two rails are a distance l apart and are connected to each other at the bottom of the incline by a zero-resistance wire. The rails are inclined at an angle θ, and there is a uniform vertical magnetic field $\vec{\mathbf{B}}$ throughout the region. (a) Show that the bar acquires a terminal velocity given by $v = \frac{mgR\sin\theta}{B^2l^2\cos^2\theta}$. (b) Calculate the work per unit time done by the force of gravity. (c) Compare this with the power dissipated in the Joule heating of the bar. (d) What would happen if $\vec{\mathbf{B}}$ were reversed?



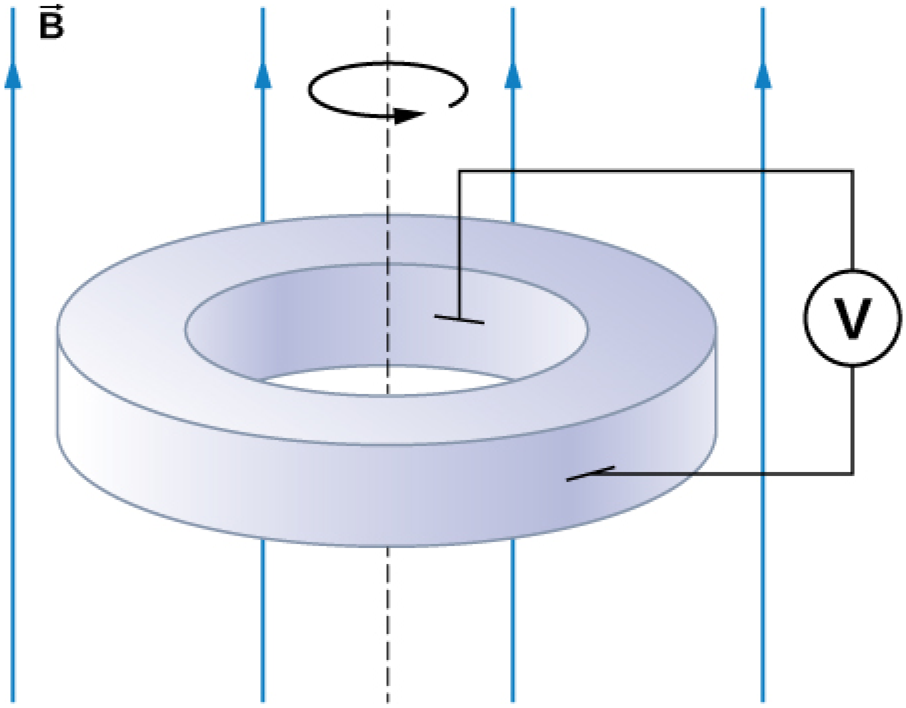
Solution:

a. $v = \frac{mgR\sin\theta}{B^2l^2\cos^2\theta}$; b. mgv sin θ; c. $mc\Delta T$; d. current would reverse direction but bar would still slide at the same speed

Exercise:

Problem:

The accompanying figure shows a metal disk of inner radius r_1 and other radius r_2 rotating at an angular velocity $\vec{\omega}$ while in a uniform magnetic field directed parallel to the rotational axis. The brush leads of a voltmeter are connected to the dark's inner and outer surfaces as shown. What is the reading of the voltmeter?



Exercise:

Problem:

A long solenoid with 10 turns per centimeter is placed inside a copper ring such that both objects have the same central axis. The radius of the ring is 10.0 cm, and the radius of the solenoid is 5.0 cm. (a) What is the emf induced in the ring when the current I through the solenoid is 5.0 A and changing at a rate of 100 A/s? (b) What is the emf induced in the ring when $I = 2.0$ A and $dI/dt = 100$ A/s? (c) What is the electric field inside the ring for these two cases? (d) Suppose the ring is moved so that its central axis and the central axis of the solenoid are still parallel but no longer coincide. (You should assume that the solenoid is still inside the ring.) Now what is the emf induced in the ring? (e) Can you calculate the electric field in the ring as you did in part (c)?

Solution:

a.

$$B = \mu_0 nI,\ \Phi_\mathrm{m} = BA = \mu_0 nIA,$$
$$\varepsilon = 9.9 \times 10^{-4}\ \mathrm{V};$$

b. 9.9×10^{-4} V;

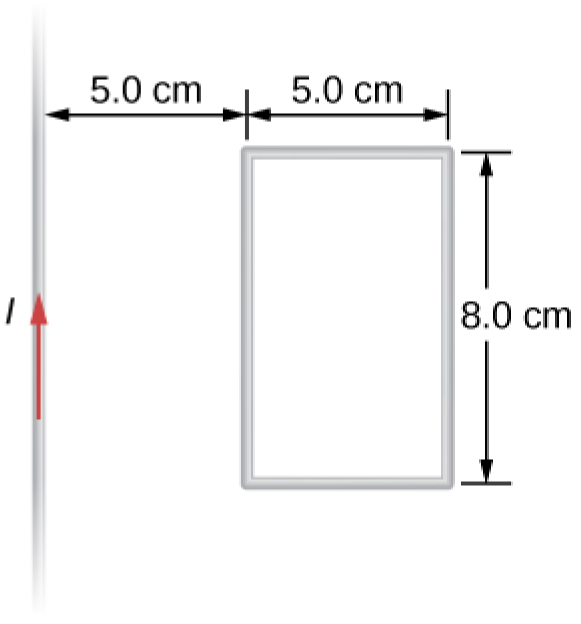
c. $\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{l}} = \varepsilon, \Rightarrow E = 1.6 \times 10^{-3}\mathrm{V/m}$; d. 9.9×10^{-4} V;

e. no, because there is no cylindrical symmetry

Exercise:

Problem:

The current in the long, straight wire shown in the accompanying figure is given by $I = I_0 \sin \omega t$, where $I_0 = 15$ A and $\omega = 120\pi\ \mathrm{rad/s}$. What is the current induced in the rectangular loop at (a) $t = 0$ and (b) $t = 2.1 \times 10^{-3}$s? The resistance of the loop is $2.0\ \Omega$.



Exercise:

Problem:

A 500-turn coil with a 0.250-m^2 area is spun in Earth's $5.00 \times 10^{-5}\text{T}$ magnetic field, producing a 12.0-kV maximum emf. (a) At what angular velocity must the coil be spun? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

a. $1.92 \times 10^6\,\text{rad/s} = 1.83 \times 10^7\,\text{rpm}$; b. This angular velocity is unreasonably high, higher than can be obtained for any mechanical system. c. The assumption that a voltage as great as 12.0 kV could be obtained is unreasonable.

Exercise:

Problem:

A circular loop of wire of radius 10 cm is mounted on a vertical shaft and rotated at a frequency of 5 cycles per second in a region of uniform magnetic field of $2 \times 10^{-4} T$ perpendicular to the axis of rotation. (a) Find an expression for the time-dependent flux through the ring (b) Determine the time-dependent current through the ring if it has a resistance of $10\,\Omega$.

Exercise:

Problem:

A long solenoid of radius a with n turns per unit length is carrying a time-dependent current $I(t) = I_0 \sin \omega t$ where I_0 and ω are constants. The solenoid is surrounded by a wire of resistance R that has two circular loops of radius b with $b > a$. Find the magnitude and direction of current induced in the outer loops at time $t = 0$.

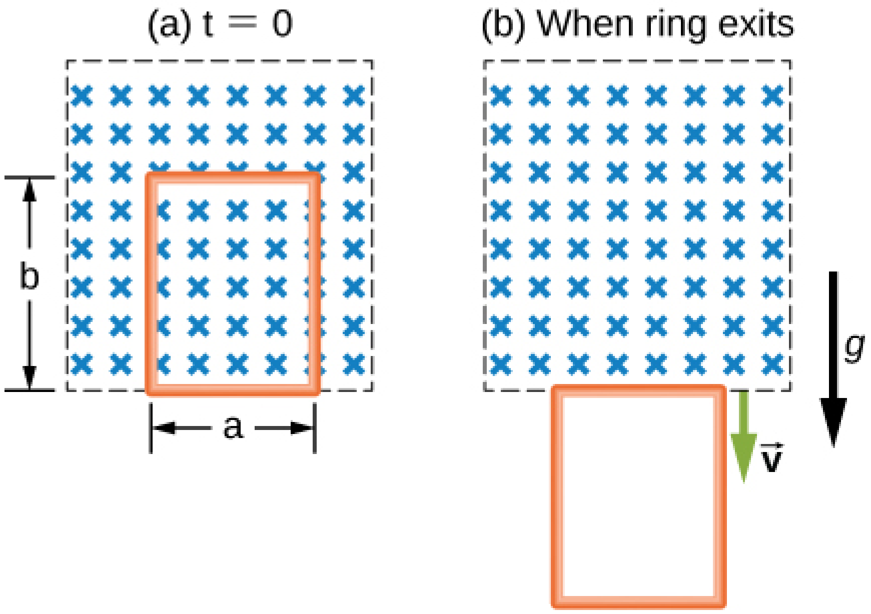
Solution:

$\frac{2\mu_0 \pi a^2 I_0 n\omega}{R}$

Exercise:

Problem:

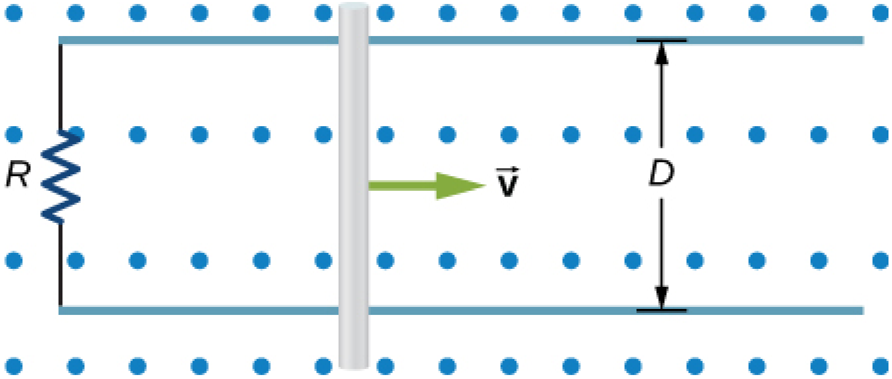
A rectangular copper loop of mass 100 g and resistance $0.2\ \Omega$ is in a region of uniform magnetic field that is perpendicular to the area enclosed by the ring and horizontal to Earth's surface (see below). The loop is let go from rest when it is at the edge of the nonzero magnetic field region. (a) Find an expression for the speed when the loop just exits the region of uniform magnetic field. (b) If it was let go at $t = 0$, what is the time when it exits the region of magnetic field for the following values: $a = 25\ \text{cm}$, $b = 50\ \text{cm}$, $B = 3\ \text{T}$, $g = 9.8\ \text{m/s}^2$? Assume that the magnetic field of the induced current is negligible compared to 3 T.



Exercise:

Problem:

A metal bar of mass *m* slides without friction over two rails a distance *D* apart in the region that has a uniform magnetic field of magnitude B_0 and direction perpendicular to the rails (see below). The two rails are connected at one end to a resistor whose resistance is much larger than the resistance of the rails and the bar. The bar is given an initial speed of v_0. It is found to slow down. How far does the bar go before coming to rest? Assume that the magnetic field of the induced current is negligible compared to B_0.



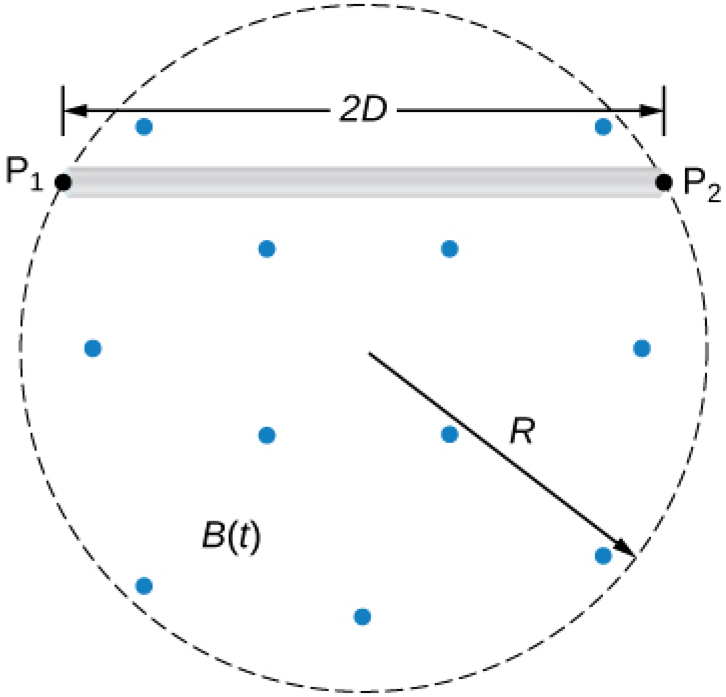
Solution:

$$\frac{mRv_o}{B^2D^2}$$

Exercise:

Problem:

A time-dependent uniform magnetic field of magnitude *B*(*t*) is confined in a cylindrical region of radius *R*. A conducting rod of length 2*D* is placed in the region, as shown below. Show that the emf between the ends of the rod is given by $\frac{dB}{dt}D\sqrt{R^2 - D^2}$. (*Hint:* To find the emf between the ends, we need to integrate the electric field from one end to the other. To find the electric field, use Faraday’s law as “Ampère’s law for *E*.”)


2D
P$_1$
P$_2$
R
B(t)

Introduction class="introduction"

A smartphone charging mat contains a coil that receives alternating current, or current that is constantly increasing and decreasing. The varying current induces an emf in the smartphone, which charges its battery. Note that the black box containing the electrical plug also contains a transformer (discussed in Alternating-Current Circuits) that modifies the current from the outlet to suit the needs of the smartphone. (credit: modification of work by “LG”/Flickr)

In Electromagnetic Induction, we discussed how a time-varying magnetic flux induces an emf in a circuit. In many of our calculations, this flux was due to an applied time-dependent magnetic field. The reverse of this phenomenon also occurs: The current flowing in a circuit produces its own magnetic field.

In Electric Charges and Fields, we saw that induction is the process by which an emf is induced by changing electric flux and separation of a dipole. So far, we have discussed some examples of induction, although some of these applications are more effective than others. The smartphone charging mat in the chapter opener photo also works by induction. Is there a useful physical quantity related to how “effective” a given device is? The answer is yes, and that physical quantity is *inductance*. In this chapter, we look at the applications of inductance in electronic devices and how inductors are used in circuits.

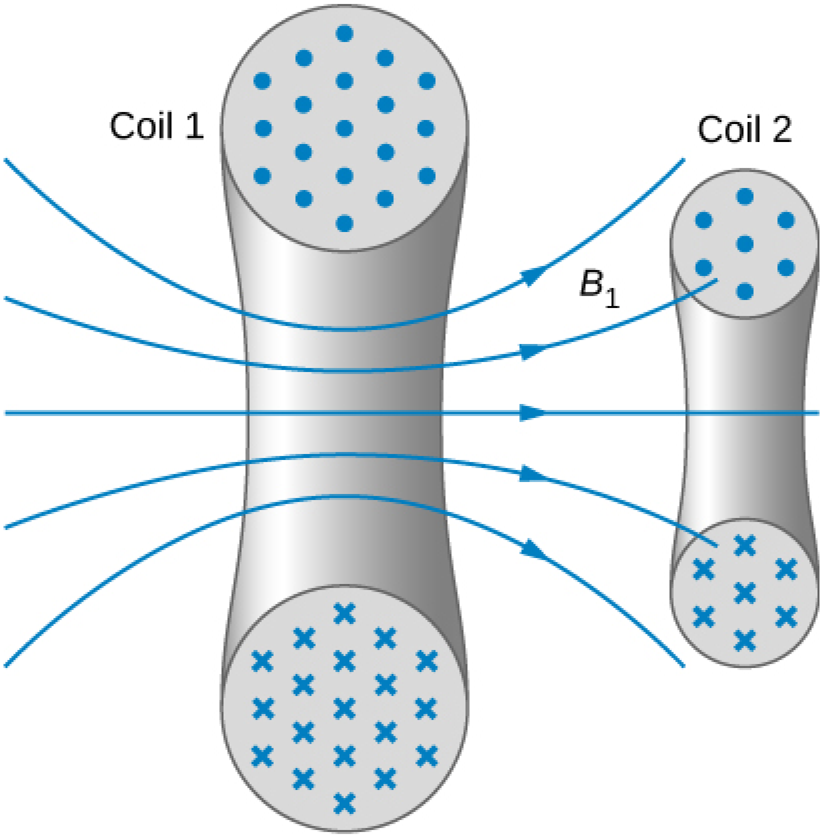
Mutual Inductance

By the end of this section, you will be able to:

- Correlate two nearby circuits that carry time-varying currents with the emf induced in each circuit
- Describe examples in which mutual inductance may or may not be desirable

Inductance is the property of a device that tells us how effectively it induces an emf in another device. In other words, it is a physical quantity that expresses the effectiveness of a given device.

When two circuits carrying time-varying currents are close to one another, the magnetic flux through each circuit varies because of the changing current I in the other circuit. Consequently, an emf is induced in each circuit by the changing current in the other. This type of emf is therefore called a *mutually induced emf*, and the phenomenon that occurs is known as **mutual inductance (*M*)**. As an example, let’s consider two tightly wound coils ([link]). Coils 1 and 2 have N_1 and N_2 turns and carry currents I_1 and I_2, respectively. The flux through a single turn of coil 2 produced by the magnetic field of the current in coil 1 is Φ_{21}, whereas the flux through a single turn of coil 1 due to the magnetic field of I_2 is Φ_{12}.



Some of the magnetic field lines produced by the current in coil 1 pass through coil 2.

The mutual inductance M_{21} of coil 2 with respect to coil 1 is the ratio of the flux through the N_2 turns of coil 2 produced by the magnetic field of the current in coil 1, divided by that current, that is,

Equation:

$$M_{21} = \frac{N_2 \Phi_{21}}{I_1}.$$

Similarly, the mutual inductance of coil 1 with respect to coil 2 is

Equation:

$$M_{12} = \frac{N_1 \Phi_{12}}{I_2}.$$

Like capacitance, mutual inductance is a geometric quantity. It depends on the shapes and relative positions of the two coils, and it is independent of the currents in the coils. The SI unit for mutual inductance *M* is called the **henry (H)** in honor of Joseph Henry (1799–1878), an American scientist who discovered induced emf independently of Faraday. Thus, we have $1\ \text{H} = 1\ \text{V} \cdot \text{s}/\text{A}$. From [link] and [link], we can show that $M_{21} = M_{12}$, so we usually drop the subscripts associated with mutual inductance and write

Note:
Equation:

$$M = \frac{N_2 \Phi_{21}}{I_1} = \frac{N_1 \Phi_{12}}{I_2}.$$

The emf developed in either coil is found by combining Faraday's law and the definition of mutual inductance. Since $N_2 \Phi_{21}$ is the total flux through coil 2 due to I_1, we obtain

Equation:

$$\varepsilon_2 = -\frac{d}{dt}(N_2 \Phi_{21}) = -\frac{d}{dt}(M I_1) = -M \frac{dI_1}{dt}$$

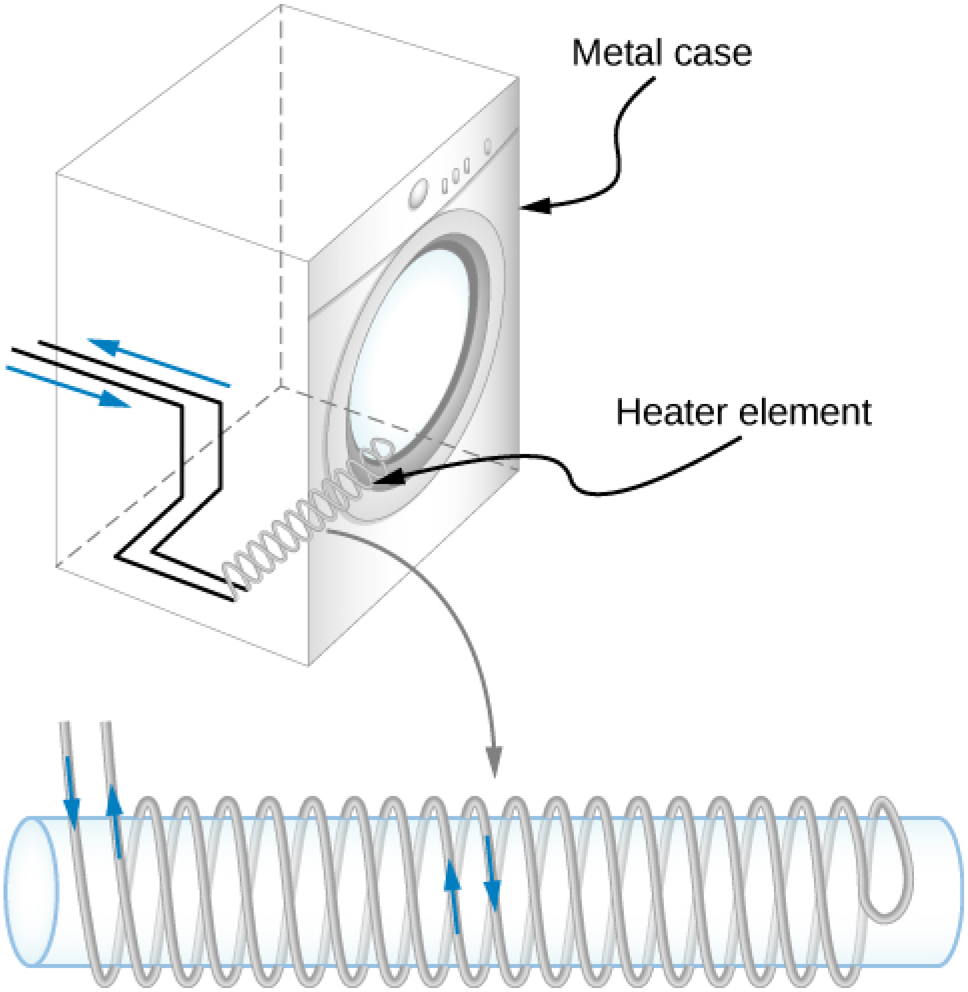
where we have used the fact that *M* is a time-independent constant because the geometry is time-independent. Similarly, we have

Note:
Equation:

$$\varepsilon_1 = -M\frac{dI_2}{dt}.$$

In [link], we can see the significance of the earlier description of mutual inductance (*M*) as a geometric quantity. The value of *M* neatly encapsulates the physical properties of circuit elements and allows us to separate the physical layout of the circuit from the dynamic quantities, such as the emf and the current. [link] defines the mutual inductance in terms of properties in the circuit, whereas the previous definition of mutual inductance in [link] is defined in terms of the magnetic flux experienced, regardless of circuit elements. You should be careful when using [link] and [link] because ε_1 and ε_2 do not necessarily represent the total emfs in the respective coils. Each coil can also have an emf induced in it because of its *self-inductance* (self-inductance will be discussed in more detail in a later section).

A large mutual inductance *M* may or may not be desirable. We want a transformer to have a large mutual inductance. But an appliance, such as an electric clothes dryer, can induce a dangerous emf on its metal case if the mutual inductance between its coils and the case is large. One way to reduce mutual inductance is to counter-wind coils to cancel the magnetic field produced ([link]).



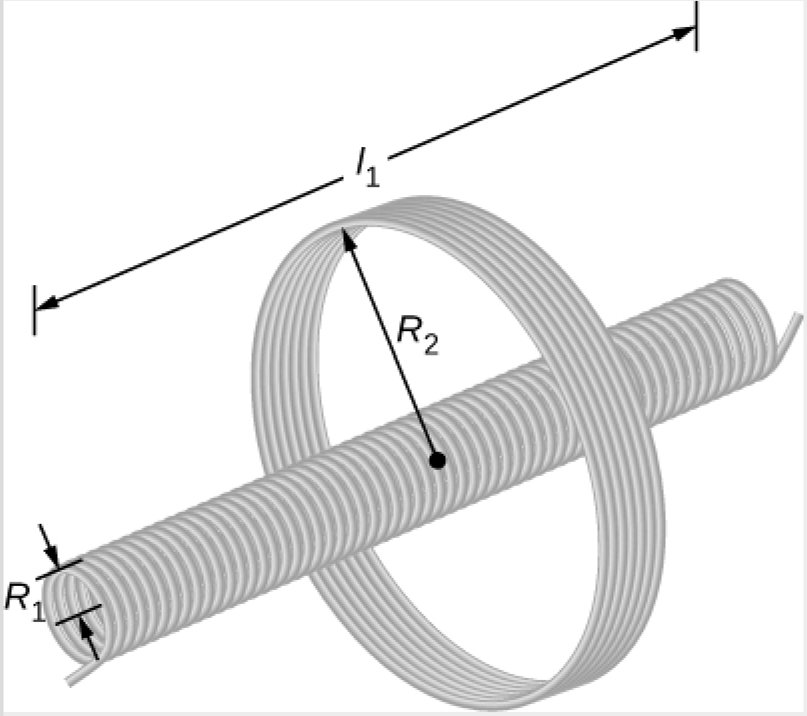
The heating coils of an electric clothes dryer can be counter-wound so that their magnetic fields cancel one another, greatly reducing the mutual inductance with the case of the dryer.

Digital signal processing is another example in which mutual inductance is reduced by counter-winding coils. The rapid on/off emf representing 1s and 0s in a digital circuit creates a complex time-dependent magnetic field. An emf can be generated in neighboring conductors. If that conductor is also carrying a digital signal, the induced emf may be large enough to switch 1s and 0s, with consequences ranging from inconvenient to disastrous.

Example:

Mutual Inductance

[link] shows a coil of N_2 turns and radius R_2 surrounding a long solenoid of length l_1, radius R_1, and N_1 turns. (a) What is the mutual inductance of the two coils? (b) If $N_1 = 500\ \text{turns}$, $N_2 = 10\ \text{turns}$, $R_1 = 3.10\ \text{cm}$, $l_1 = 75.0\ \text{cm}$, and the current in the solenoid is changing at a rate of 200 A/s, what is the emf induced in the surrounding coil?



A solenoid surrounded by a coil.

Strategy

There is no magnetic field outside the solenoid, and the field inside has magnitude $B_1 = \mu_0(N_1/l_1)I_1$ and is directed parallel to the solenoid's axis. We can use this magnetic field to find the magnetic flux through the surrounding coil and then use this flux to calculate the mutual inductance for part (a), using [link]. We solve part (b) by calculating the mutual inductance from the given quantities and using [link] to calculate the induced emf.

Solution

a. The magnetic flux Φ_{21} through the surrounding coil is
Equation:

$$\Phi_{21} = B_1 \pi R_1^2 = \frac{\mu_0 N_1 I_1}{l_1} \pi R_1^2.$$

Now from [link], the mutual inductance is
Equation:

$$M = \frac{N_2 \Phi_{21}}{I_1} = \frac{N_2}{I_1} \frac{\mu_0 N_1 I_1}{l_1} \pi R_1^2 = \frac{\mu_0 N_1 N_2 \pi R_1^2}{l_1}.$$

b. Using the previous expression and the given values, the mutual inductance is
Equation:

$$\begin{aligned} M &= \frac{(4\pi \times 10^{-7}\,\text{T}\cdot\text{m/A})(500)(10)\pi(0.0310\,\text{m})^2}{0.750\,\text{m}} \\ &= 2.53 \times 10^{-5}\,\text{H}. \end{aligned}$$

Thus, from [link], the emf induced in the surrounding coil is
Equation:

$$\begin{aligned} \varepsilon_2 &= -M \frac{dI_1}{dt} = -(2.53 \times 10^{-5}\,\text{H})(200\,\text{A/s}) \\ &= -5.06 \times 10^{-3}\,\text{V}. \end{aligned}$$

Significance

Notice that M in part (a) is independent of the radius R_2 of the surrounding coil because the solenoid's magnetic field is confined to its interior. In principle, we can also calculate M by finding the magnetic flux through the solenoid produced by the current in the surrounding coil. This approach is much more difficult because Φ_{12} is so complicated. However, since $M_{12} = M_{21}$, we do know the result of this calculation.

Note:
Exercise:

Problem:

Check Your Understanding A current $I(t) = (5.0\text{ A})\sin((120\pi\text{ rad/s})t)$ flows through the solenoid of part (b) of [link]. What is the maximum emf induced in the surrounding coil?

Solution:

$4.77 \times 10^{-2}\text{ V}$

Summary

- Inductance is the property of a device that expresses how effectively it induces an emf in another device.
- Mutual inductance is the effect of two devices inducing emfs in each other.
- A change in current dI_1/dt in one circuit induces an emf (ε_2) in the second:
  Equation:
  $$\varepsilon_2 = -M\frac{dI1}{dt},$$
  where M is defined to be the mutual inductance between the two circuits and the minus sign is due to Lenz's law.
- Symmetrically, a change in current dI_2/dt through the second circuit induces an emf (ε_1) in the first:
  Equation:
  $$\varepsilon_1 = -M\frac{dI_2}{dt},$$

where M is the same mutual inductance as in the reverse process.

Conceptual Questions

Exercise:

Problem:

Show that $N\Phi_{\mathrm{m}}/I$ and $\varepsilon/(dI/dt)$, which are both expressions for self-inductance, have the same units.

Solution:

$\frac{\mathrm{Wb}}{\mathrm{A}} = \frac{\mathrm{T{\cdot}m^2}}{\mathrm{A}} = \frac{\mathrm{V{\cdot}s}}{\mathrm{A}} = \frac{\mathrm{V}}{\mathrm{A/s}}$

Exercise:

Problem:

A 10-H inductor carries a current of 20 A. Describe how a 50-V emf can be induced across it.

Exercise:

Problem:

The ignition circuit of an automobile is powered by a 12-V battery. How are we able to generate large voltages with this power source?

Solution:

The induced current from the 12-V battery goes through an inductor, generating a large voltage.

Exercise:

Problem:

When the current through a large inductor is interrupted with a switch, an arc appears across the open terminals of the switch. Explain.

Problems

Exercise:

Problem:

When the current in one coil changes at a rate of 5.6 A/s, an emf of 6.3×10^{-3} V is induced in a second, nearby coil. What is the mutual inductance of the two coils?

Exercise:

Problem:

An emf of 9.7×10^{-3} V is induced in a coil while the current in a nearby coil is decreasing at a rate of 2.7 A/s. What is the mutual inductance of the two coils?

Solution:

$M = 3.6 \times 10^{-3}$ H

Exercise:

Problem:

Two coils close to each other have a mutual inductance of 32 mH. If the current in one coil decays according to $I = I_0 e^{-\alpha t}$, where $I_0 = 5.0$ A and $\alpha = 2.0 \times 10^3$ s^{-1}, what is the emf induced in the second coil immediately after the current starts to decay? At $t = 1.0 \times 10^{-3}$ s?

Exercise:

Problem:

A coil of 40 turns is wrapped around a long solenoid of cross-sectional area $7.5 \times 10^{-3}\ \text{m}^2$. The solenoid is 0.50 m long and has 500 turns. (a) What is the mutual inductance of this system? (b) The outer coil is replaced by a coil of 40 turns whose radius is three times that of the solenoid. What is the mutual inductance of this configuration?

Solution:

a. $3.8 \times 10^{-4}\ \text{H}$; b. $3.8 \times 10^{-4}\ \text{H}$

Exercise:

Problem:

A 600-turn solenoid is 0.55 m long and 4.2 cm in diameter. Inside the solenoid, a small $(1.1\ \text{cm} \times 1.4\ \text{cm})$, single-turn rectangular coil is fixed in place with its face perpendicular to the long axis of the solenoid. What is the mutual inductance of this system?

Exercise:

Problem:

A toroidal coil has a mean radius of 16 cm and a cross-sectional area of $0.25\ \text{cm}^2$; it is wound uniformly with 1000 turns. A second toroidal coil of 750 turns is wound uniformly over the first coil. Ignoring the variation of the magnetic field within a toroid, determine the mutual inductance of the two coils.

Solution:

$M_{21} = 2.3 \times 10^{-5}\ \text{H}$

Exercise:

Problem:

A solenoid of N_1 turns has length l_1 and radius R_1, and a second smaller solenoid of N_2 turns has length l_2 and radius R_2. The smaller solenoid is placed completely inside the larger solenoid so that their long axes coincide. What is the mutual inductance of the two solenoids?

Glossary

henry (H)	unit of inductance, $1\ \text{H} = 1\ \Omega \cdot \text{s}$; it is also expressed as a volt second per ampere
inductance	property of a device that tells how effectively it induces an emf in another device
mutual inductance	geometric quantity that expresses how effective two devices are at inducing emfs in one another

Self-Inductance and Inductors

By the end of this section, you will be able to:

- Correlate the rate of change of current to the induced emf created by that current in the same circuit
- Derive the self-inductance for a cylindrical solenoid
- Derive the self-inductance for a rectangular toroid

Mutual inductance arises when a current in one circuit produces a changing magnetic field that induces an emf in another circuit. But can the magnetic field affect the current in the original circuit that produced the field? The answer is yes, and this is the phenomenon called *self-inductance*.

Inductors

[link] shows some of the magnetic field lines due to the current in a circular loop of wire. If the current is constant, the magnetic flux through the loop is also constant. However, if the current I were to vary with time—say, immediately after switch S is closed—then the magnetic flux Φ_{m} would correspondingly change. Then Faraday's law tells us that an emf ε would be induced in the circuit, where

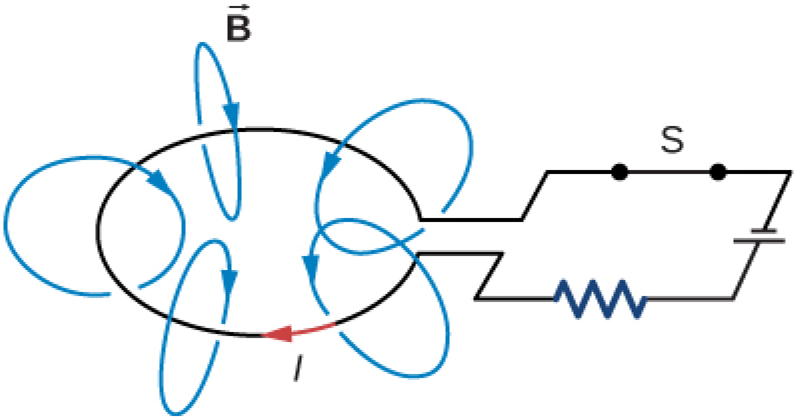
Equation:

$$\varepsilon = -\frac{d\Phi_{\mathrm{m}}}{dt}.$$

Since the magnetic field due to a current-carrying wire is directly proportional to the current, the flux due to this field is also proportional to the current; that is,

Equation:

$$\Phi_{\mathrm{m}} \propto I.$$



A magnetic field is produced by the current I in the loop. If I were to vary with time, the magnetic flux through the loop would also vary and an emf would be induced in the loop.

This can also be written as
Equation:

$$\Phi_{\rm m} = LI$$

where the constant of proportionality L is known as the **self-inductance** of the wire loop. If the loop has N turns, this equation becomes

Note:
Equation:

$$N\Phi_{\rm m} = LI.$$

By convention, the positive sense of the normal to the loop is related to the current by the right-hand rule, so in [link], the normal points downward.

With this convention, Φ_{m} is positive in [link], so L *always has a positive value*.

For a loop with N turns, $\varepsilon = -N d\Phi_{\text{m}}/dt$, so the induced emf may be written in terms of the self-inductance as

Note:
Equation:

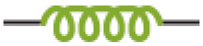
$$\varepsilon = -L\frac{dI}{dt}.$$

When using this equation to determine L, it is easiest to ignore the signs of ε and dI/dt, and calculate L as
Equation:

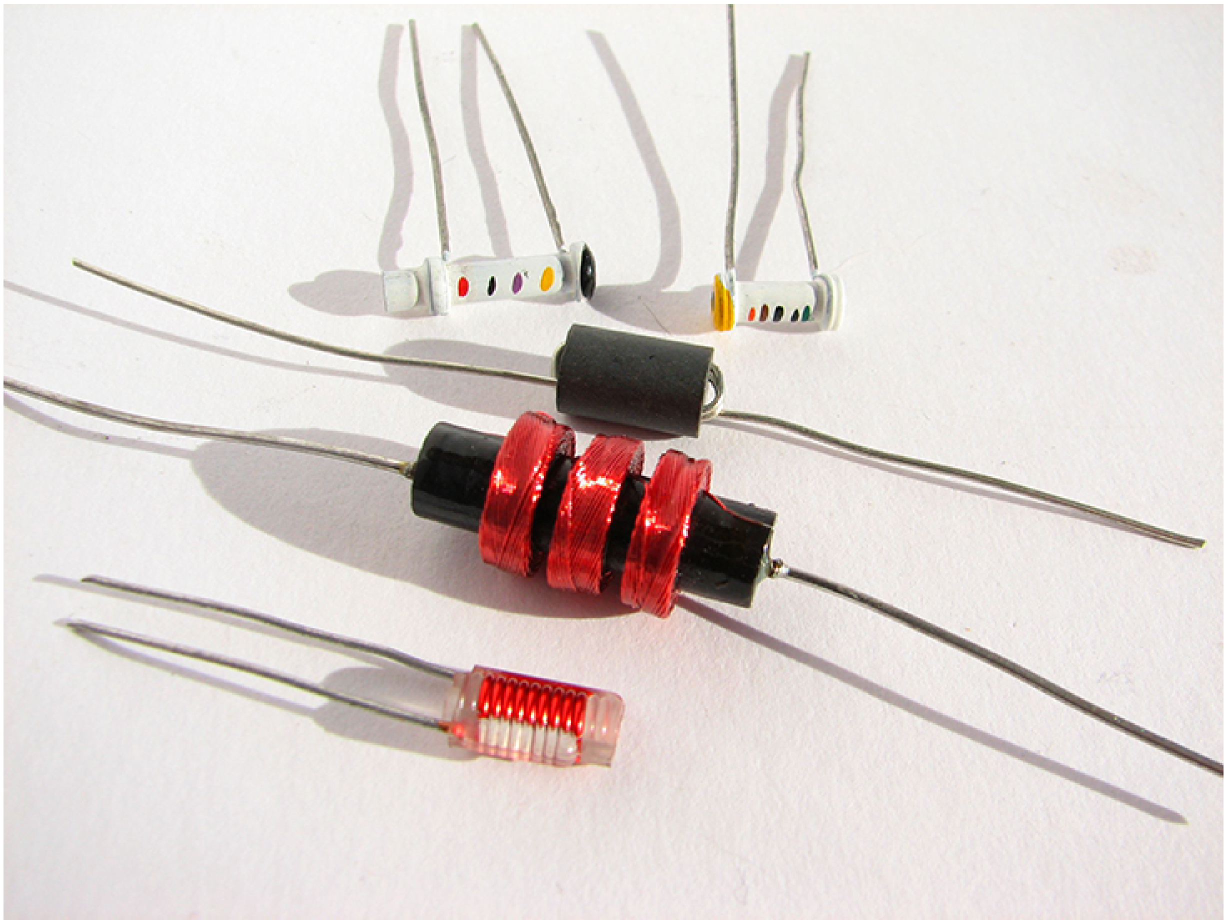
$$L = \frac{|\varepsilon|}{|dI/dt|}.$$

Since self-inductance is associated with the magnetic field produced by a current, any configuration of conductors possesses self-inductance. For example, besides the wire loop, a long, straight wire has self-inductance, as does a coaxial cable. A coaxial cable is most commonly used by the cable television industry and may also be found connecting to your cable modem. Coaxial cables are used due to their ability to transmit electrical signals with minimal distortions. Coaxial cables have two long cylindrical conductors that possess current and a self-inductance that may have undesirable effects.

A circuit element used to provide self-inductance is known as an **inductor**. It is represented by the symbol shown in [link], which resembles a coil of wire, the basic form of the inductor. [link] shows several types of inductors commonly used in circuits.

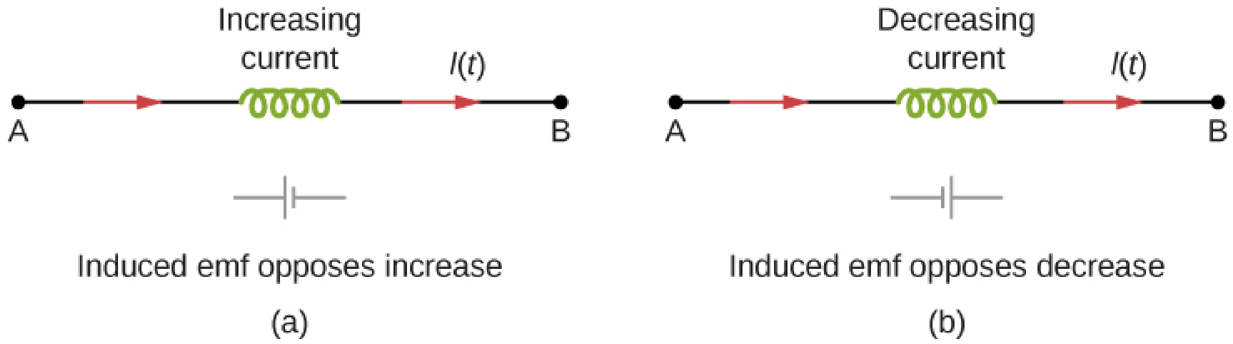


Symbol used to represent an inductor in a circuit.



A variety of inductors. Whether they are encapsulated like the top three shown or wound around in a coil like the bottom-most one, each is simply a relatively long coil of wire. (credit: Windell Oskay)

In accordance with Lenz's law, the negative sign in [link] indicates that the induced emf across an inductor always has a polarity that *opposes* the change in the current. For example, if the current flowing from *A* to *B* in [link](a) were increasing, the induced emf (represented by the imaginary battery) would have the polarity shown in order to oppose the increase. If the current from *A* to *B* were decreasing, then the induced emf would have the opposite polarity, again to oppose the change in current ([link](b)). Finally, if the current through the inductor were constant, no emf would be induced in the coil.


Increasing
current
I(t)
A
B
Induced emf opposes increase
(a)
Decreasing
current
I(t)
A
B
Induced emf opposes decrease
(b)

The induced emf across an inductor always acts to oppose the change in the current. This can be visualized as an imaginary battery causing current to flow to oppose the change in (a) and reinforce the change in (b).

One common application of inductance is to allow traffic signals to sense when vehicles are waiting at a street intersection. An electrical circuit with an inductor is placed in the road underneath the location where a waiting car will stop. The body of the car increases the inductance and the circuit changes, sending a signal to the traffic lights to change colors. Similarly, metal detectors used for airport security employ the same technique. A coil or inductor in the metal detector frame acts as both a transmitter and a receiver. The pulsed signal from the transmitter coil induces a signal in the receiver. The self-inductance of the circuit is affected by any metal object in the path ([link]). Metal detectors can be adjusted for sensitivity and can also sense the presence of metal on a person.

The familiar security gate at an airport not only detects metals, but can also indicate their approximate height above the floor. (credit: “Alexbuirds”/Wikimedia Commons)

Large induced voltages are found in camera flashes. Camera flashes use a battery, two inductors that function as a transformer, and a switching system or *oscillator* to induce large voltages. Recall from Oscillations on oscillations that “oscillation” is defined as the fluctuation of a quantity, or repeated regular fluctuations of a quantity, between two extreme values around an average value. Also recall (from Electromagnetic Induction on electromagnetic induction) that we need a changing magnetic field, brought about by a changing current, to induce a voltage in another coil. The oscillator system does this many times as the battery voltage is boosted to over 1000 volts. (You may hear the high-pitched whine from the transformer as the capacitor is being charged.) A capacitor stores the high voltage for later use in powering the flash.

Example:

Self-Inductance of a Coil

An induced emf of 2.0 V is measured across a coil of 50 closely wound turns while the current through it increases uniformly from 0.0 to 5.0 A in 0.10 s. (a) What is the self-inductance of the coil? (b) With the current at 5.0 A, what is the flux through each turn of the coil?

Strategy

Both parts of this problem give all the information needed to solve for the self-inductance in part (a) or the flux through each turn of the coil in part (b). The equations needed are [link] for part (a) and [link] for part (b).

Solution

a. Ignoring the negative sign and using magnitudes, we have, from [link],
   Equation:

$$L=\frac{\varepsilon}{dI/dt}=\frac{2.0\ \text{V}}{5.0\ \text{A}/0.10\ \text{s}}=4.0\times 10^{-2}\ \text{H}.$$

b. From [link], the flux is given in terms of the current by $\Phi_{\text{m}}=LI/N$, so
   Equation:

$$\Phi_{\text{m}}=\frac{(4.0\times 10^{-2}\ \text{H})(5.0\ \text{A})}{50\ \text{turns}}=4.0\times 10^{-3}\ \text{Wb}.$$

Significance

The self-inductance and flux calculated in parts (a) and (b) are typical values for coils found in contemporary devices. If the current is not changing over time, the flux is not changing in time, so no emf is induced.

Note:
Exercise:

Problem:

Check Your Understanding Current flows through the inductor in [link] from *B* to *A* instead of from *A* to *B* as shown. Is the current increasing or decreasing in order to produce the emf given in diagram (a)? In diagram (b)?

Solution:

a. decreasing; b. increasing; Since the current flows in the opposite direction of the diagram, in order to get a positive emf on the left-hand side of diagram (a), we need to decrease the current to the left, which creates a reinforced emf where the positive end is on the left-hand side. To get a positive emf on the right-hand side of diagram (b), we need to increase the current to the left, which creates a reinforced emf where the positive end is on the right-hand side.

Note:
Exercise:

Problem:

Check Your Understanding A changing current induces an emf of 10 V across a 0.25-H inductor. What is the rate at which the current is changing?

Solution:

40 A/s

A good approach for calculating the self-inductance of an inductor consists of the following steps:

Note:
Problem-Solving Strategy: Self-Inductance

1. Assume a current I is flowing through the inductor.
2. Determine the magnetic field $\vec{\ }$ produced by the current. If there is appropriate symmetry, you may be able to do this with Ampère’s law.
3. Obtain the magnetic flux, Φ_{m}.
4. With the flux known, the self-inductance can be found from [link], $L = N\Phi_{\mathrm{m}}/I$.

To demonstrate this procedure, we now calculate the self-inductances of two inductors.

Cylindrical Solenoid

Consider a long, cylindrical solenoid with length l, cross-sectional area A, and N turns of wire. We assume that the length of the solenoid is so much larger than its diameter that we can take the magnetic field to be $B = \mu_0 nI$ throughout the interior of the solenoid, that is, we ignore end effects in the solenoid. With a current I flowing through the coils, the magnetic field produced within the solenoid is

Equation:

$$B = \mu_0 \frac{N}{l} I,$$

so the magnetic flux through one turn is

Equation:

$$\Phi_{\mathrm{m}} = BA = \frac{\mu_0 NA}{l} I.$$

Using [link], we find for the self-inductance of the solenoid,

Note:
Equation:

$$L_{\text{solenoid}} = \frac{N\Phi_{\mathrm{m}}}{I} = \frac{\mu_0 N^2 A}{l}.$$

If $n = N/l$ is the number of turns per unit length of the solenoid, we may write [link] as

Equation:

$$L = \mu_0 {\frac{N}{l}}^{2} Al = \mu_0 n^2 Al = \mu_0 n^2 (V),$$

where $V = Al$ is the volume of the solenoid. Notice that *the self-inductance of a long solenoid depends only on its physical properties* (such as the number of turns of wire per unit length and the volume), and not on the magnetic field or the current. This is true for inductors in general.

Rectangular Toroid

A toroid with a rectangular cross-section is shown in [link]. The inner and outer radii of the toroid are R_1 and $R_2,$ and h is the height of the toroid. Applying Ampère's law in the same manner as we did in [link] for a toroid with a circular cross-section, we find the magnetic field inside a rectangular toroid is also given by

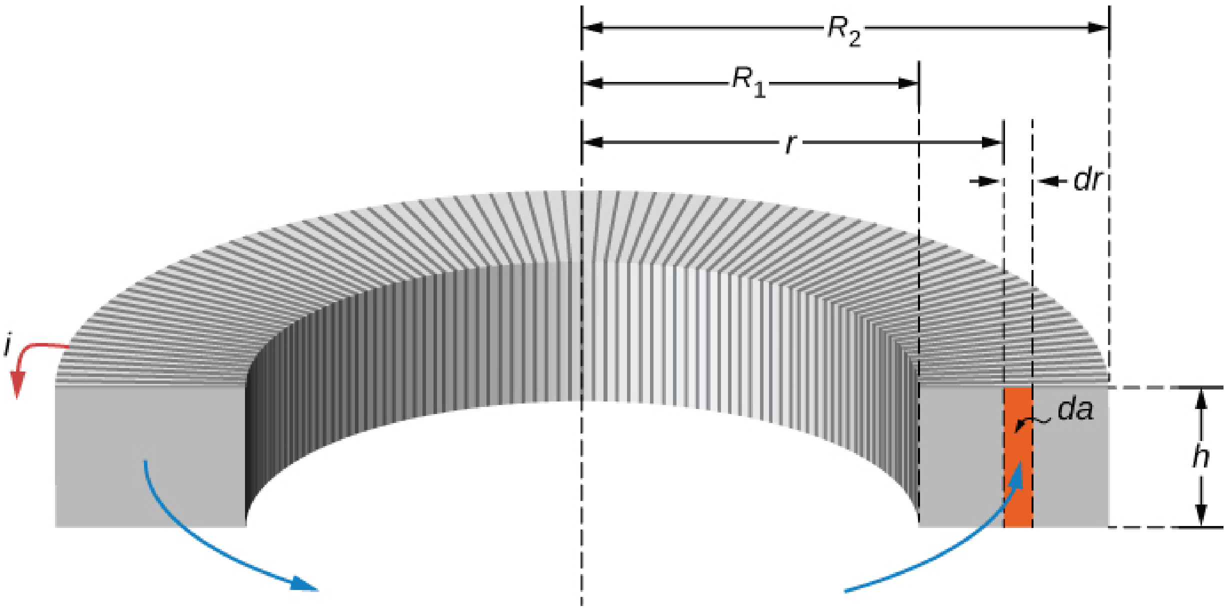
Equation:

$$B = \frac{\mu_0 NI}{2\pi r},$$

where r is the distance from the central axis of the toroid. Because the field changes within the toroid, we must calculate the flux by integrating over the toroid’s cross-section. Using the infinitesimal cross-sectional area element $da = h\,dr$ shown in [link], we obtain

Equation:

$$\Phi_{\text{m}} = \int B\,da = \int_{R_1}^{R_2} \frac{\mu_0 NI}{2\pi r}(hdr) = \frac{\mu_0 NhI}{2\pi} \ln \frac{R_2}{R_1}.$$



Calculating the self-inductance of a rectangular toroid.

Now from [link], we obtain for the self-inductance of a rectangular toroid

Note:
Equation:

$$L = \frac{N\Phi_{\text{m}}}{I} = \frac{\mu_0 N^2 h}{2\pi} \ln \frac{R_2}{R_1}.$$

As expected, the self-inductance is a constant determined by only the physical properties of the toroid.

Note:
Exercise:

Problem:

Check Your Understanding (a) Calculate the self-inductance of a solenoid that is tightly wound with wire of diameter 0.10 cm, has a cross-sectional area of $0.90\ \text{cm}^2$, and is 40 cm long. (b) If the current through the solenoid decreases uniformly from 10 to 0 A in 0.10 s, what is the emf induced between the ends of the solenoid?

Solution:

a. 4.5×10^{-5} H; b. 4.5×10^{-3} V

Note:
Exercise:

Problem:

Check Your Understanding (a) What is the magnetic flux through one turn of a solenoid of self-inductance 8.0×10^{-5} H when a current of 3.0 A flows through it? Assume that the solenoid has 1000 turns and is wound from wire of diameter 1.0 mm. (b) What is the cross-sectional area of the solenoid?

Solution:

a. 2.4×10^{-7} Wb; b. $6.4 \times 10^{-5}\text{ m}^2$

Summary

- Current changes in a device induce an emf in the device itself, called self-inductance,
  Equation:
  $$\varepsilon = -L\frac{dI}{dt},$$
  where L is the self-inductance of the inductor and dI/dt is the rate of change of current through it. The minus sign indicates that emf opposes the change in current, as required by Lenz’s law. The unit of self-inductance and inductance is the henry (H), where $1\text{ H} = 1\,\Omega \cdot \text{s}$.
- The self-inductance of a solenoid is
  Equation:
  $$L = \frac{\mu_0 N^2 A}{l},$$
  where N is its number of turns in the solenoid, A is its cross-sectional

area, *l* is its length, and $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{T \cdot m/A}$ is the permeability of free space.

- The self-inductance of a toroid is
  Equation:

  $$L = \frac{\mu_0 N^2 h}{2\pi} \ln \frac{R_2}{R_1},$$

  where *N* is its number of turns in the toroid, R_1 and R_2 are the inner and outer radii of the toroid, *h* is the height of the toroid, and $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{T \cdot m/A}$ is the permeability of free space.

Conceptual Questions

Exercise:

Problem:

Does self-inductance depend on the value of the magnetic flux? Does it depend on the current through the wire? Correlate your answers with the equation $N\Phi_{\mathrm{m}} = LI$.

Solution:

Self-inductance is proportional to the magnetic flux and inversely proportional to the current. However, since the magnetic flux depends on the current *I*, these effects cancel out. This means that the self-inductance does not depend on the current. If the emf is induced across an element, it does depend on how the current changes with time.

Exercise:

Problem:

Would the self-inductance of a 1.0 m long, tightly wound solenoid differ from the self-inductance per meter of an infinite, but otherwise identical, solenoid?

Exercise:

Problem:

Discuss how you might determine the self-inductance per unit length of a long, straight wire.

Solution:

Consider the ends of a wire a part of an *RL* circuit and determine the self-inductance from this circuit.

Exercise:

Problem:

The self-inductance of a coil is zero if there is no current passing through the windings. True or false?

Exercise:

Problem:

How does the self-inductance per unit length near the center of a solenoid (away from the ends) compare with its value near the end of the solenoid?

Solution:

The magnetic field will flare out at the end of the solenoid so there is less flux through the last turn than through the middle of the solenoid.

Problems

Exercise:

Problem:

An emf of 0.40 V is induced across a coil when the current through it changes uniformly from 0.10 to 0.60 A in 0.30 s. What is the self-inductance of the coil?

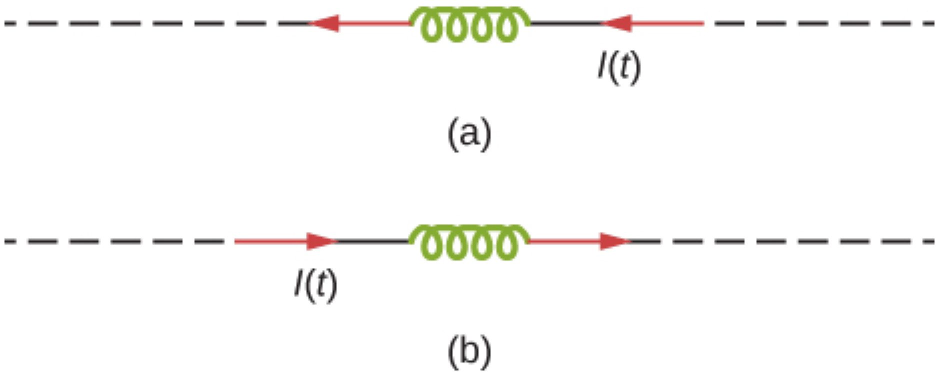
Solution:

0.24 H

Exercise:

Problem:

The current shown in part (a) below is increasing, whereas that shown in part (b) is decreasing. In each case, determine which end of the inductor is at the higher potential.



Exercise:

Problem:

What is the rate at which the current though a 0.30-H coil is changing if an emf of 0.12 V is induced across the coil?

Solution:

0.4 A/s

Exercise:

Problem:

When a camera uses a flash, a fully charged capacitor discharges through an inductor. In what time must the 0.100-A current through a 2.00-mH inductor be switched on or off to induce a 500-V emf?

Exercise:

Problem:

A coil with a self-inductance of 2.0 H carries a current that varies with time according to $I(t) = (2.0\text{ A})\sin 120\pi t$. Find an expression for the emf induced in the coil.

Solution:

$\varepsilon = 480\pi \sin(120\pi t - \pi/2)\text{ V}$

Exercise:

Problem:

A solenoid 50 cm long is wound with 500 turns of wire. The cross-sectional area of the coil is 2.0 cm^2 What is the self-inductance of the solenoid?

Exercise:

Problem:

A coil with a self-inductance of 3.0 H carries a current that decreases at a uniform rate $dI/dt = -0.050\text{ A/s}$. What is the emf induced in the coil? Describe the polarity of the induced emf.

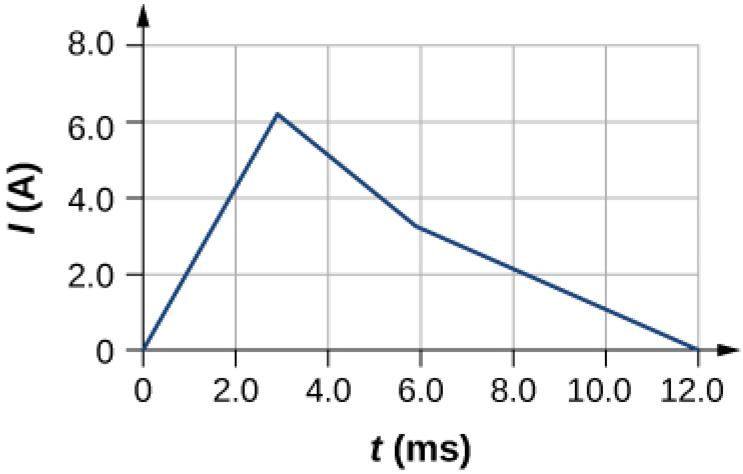
Solution:

0.15 V. This is the same polarity as the emf driving the current.

Exercise:

Problem:

The current *I(t)* through a 5.0-mH inductor varies with time, as shown below. The resistance of the inductor is $5.0\,\Omega$. Calculate the voltage across the inductor at $t=2.0\,\text{ms}$, $t=4.0\,\text{ms},$ and $t=8.0\,\text{ms}$.



Exercise:

Problem:

A long, cylindrical solenoid with 100 turns per centimeter has a radius of 1.5 cm. (a) Neglecting end effects, what is the self-inductance per unit length of the solenoid? (b) If the current through the solenoid changes at the rate 5.0 A/s, what is the emf induced per unit length?

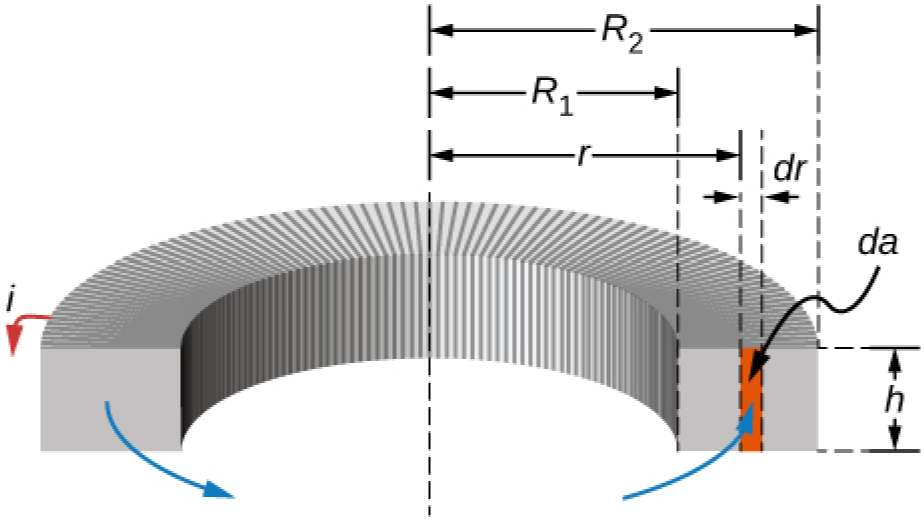
Solution:

a. 0.089 H/m; b. 0.44 V/m

Exercise:

Problem:

Suppose that a rectangular toroid has 2000 windings and a self-inductance of 0.040 H. If $h=0.10\,\text{m}$, what is the ratio of its outer radius to its inner radius?



Exercise:

Problem:

What is the self-inductance per meter of a coaxial cable whose inner radius is 0.50 mm and whose outer radius is 4.00 mm?

Solution:

$$\frac{L}{l} = 4.16 \times 10^{-7}\,\text{H/m}$$

Glossary

inductor
: part of an electrical circuit to provide self-inductance, which is symbolized by a coil of wire

self-inductance
: effect of the device inducing emf in itself

Energy in a Magnetic Field

By the end of this section, you will be able to:

- Explain how energy can be stored in a magnetic field
- Derive the equation for energy stored in a coaxial cable given the magnetic energy density

The energy of a capacitor is stored in the electric field between its plates. Similarly, an inductor has the capability to store energy, but in its magnetic field. This energy can be found by integrating the **magnetic energy density**,

Equation:

$$u_{\text{m}} = \frac{B^2}{2\mu_0}$$

over the appropriate volume. To understand where this formula comes from, let's consider the long, cylindrical solenoid of the previous section. Again using the infinite solenoid approximation, we can assume that the magnetic field is essentially constant and given by $B = \mu_0 nI$ everywhere inside the solenoid. Thus, the energy stored in a solenoid or the magnetic energy density times volume is equivalent to

Equation:

$$U = u_{\text{m}}(V) = \frac{(\mu_0 nI)^2}{2\mu_0}(Al) = \frac{1}{2}\left(\mu_0 n^2 Al\right)I^2.$$

With the substitution of [link], this becomes

Note:

Equation:

$$U = \frac{1}{2}LI^2.$$

Although derived for a special case, this equation gives the energy stored in the magnetic field of *any* inductor. We can see this by considering an arbitrary inductor through which a changing current is passing. At any instant, the magnitude of the induced emf is $\varepsilon = Ldi/dt$, where i is the induced current at that instance. Therefore, the power absorbed by the inductor is

Equation:

$$P = \varepsilon i = L\frac{di}{dt}i.$$

The total energy stored in the magnetic field when the current increases from 0 to *I* in a time interval from 0 to *t* can be determined by integrating this expression:

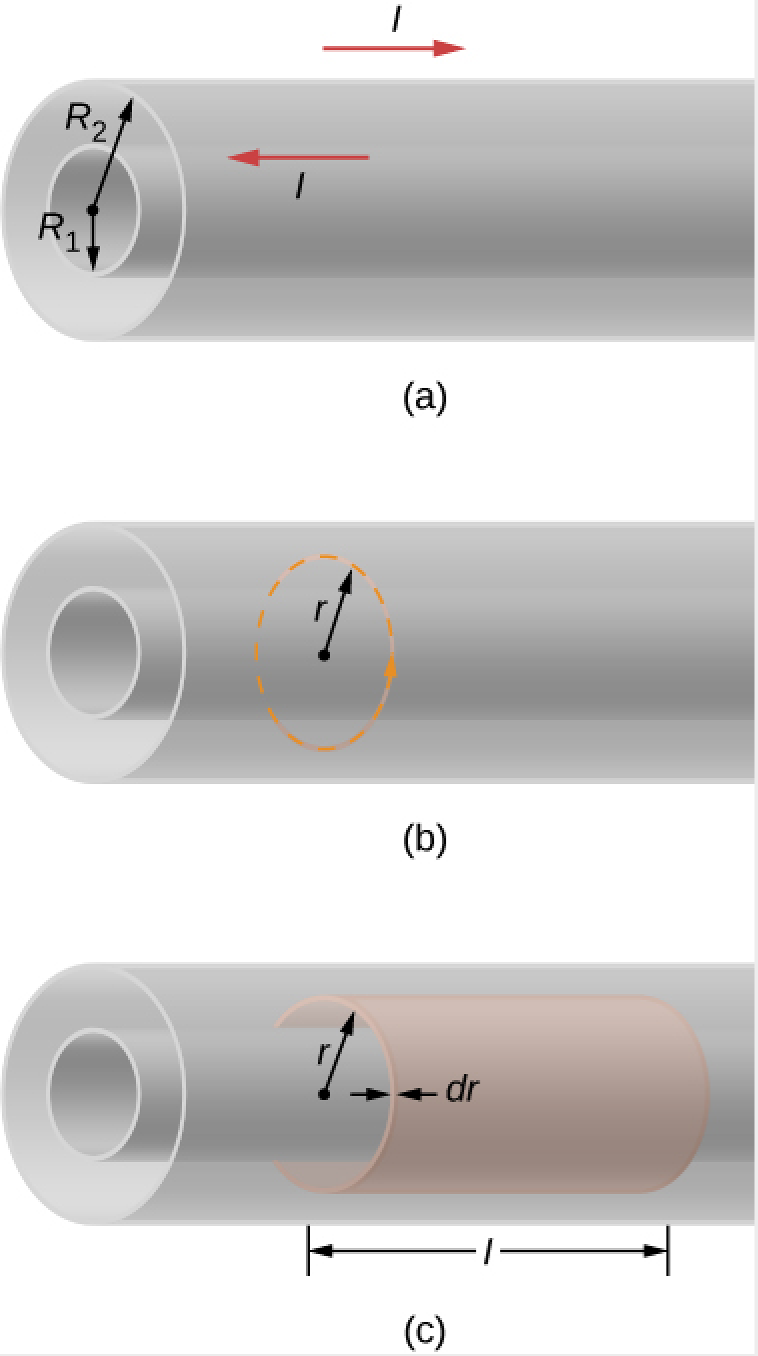
Equation:

$$U = \int_0^t P dt' = \int_0^t L\frac{di}{dt'} i dt' = L\int_0^I i di = \frac{1}{2}LI^2.$$

Example:

Self-Inductance of a Coaxial Cable

[link] shows two long, concentric cylindrical shells of radii R_1 and R_2. As discussed in Capacitance on capacitance, this configuration is a simplified representation of a coaxial cable. The capacitance per unit length of the cable has already been calculated. Now (a) determine the magnetic energy stored per unit length of the coaxial cable and (b) use this result to find the self-inductance per unit length of the cable.



(a) A coaxial cable is represented here by two hollow, concentric cylindrical conductors along which electric current flows in opposite directions. (b) The magnetic field between the conductors can be found by applying Ampère’s law to the dashed path. (c) The cylindrical shell is used to find the magnetic energy stored in a length *l* of the cable.

Strategy

The magnetic field both inside and outside the coaxial cable is determined by Ampère’s law. Based on this magnetic field, we can use [link] to calculate the energy density of the magnetic field. The magnetic energy is calculated by an integral of the magnetic energy density times the differential volume over the cylindrical shell. After the integration is carried out, we have a closed-form solution for part (a). The self-inductance per unit length is determined based on this result and [link].

Solution

a. We determine the magnetic field between the conductors by applying Ampère’s law to the dashed circular path shown in [link](b). Because of the cylindrical symmetry, $\vec{\mathbf{B}}$ is constant along the path, and

   Equation:

   $$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{l}} = B(2\pi r) = \mu_0 I.$$

   This gives us

   Equation:

   $$B = \frac{\mu_0 I}{2\pi r}.$$

   In the region outside the cable, a similar application of Ampère’s law shows that $B = 0$, since no net current crosses the area bounded by a circular path where $r > R_2$. This argument also holds when $r < R_1$; that is, in the region within the inner cylinder. All the magnetic energy of the cable is therefore stored between the two conductors. Since the energy density of the magnetic field is

   Equation:

$$u_{\mathrm{m}} = \frac{B^2}{2\mu_0}$$

the energy stored in a cylindrical shell of inner radius *r*, outer radius $r + dr$, and length *l* (see part (c) of the figure) is
Equation:

$$u_{\mathrm{m}} = \frac{\mu_0 I^2}{8\pi^2 r^2}.$$

Thus, the total energy of the magnetic field in a length *l* of the cable is
Equation:

$$U = \int_{R_1}^{R_2} dU = \int_{R_1}^{R_2} \frac{\mu_0 I^2}{8\pi^2 r^2}(2\pi r l)dr = \frac{\mu_0 I^2 l}{4\pi} \ln \frac{R_2}{R_1},$$

and the energy per unit length is $(\mu_0 I^2/4\pi)\ln(R_2/R_1)$.

b. From [link],
Equation:

$$U = \frac{1}{2} L I^2,$$

where *L* is the self-inductance of a length *l* of the coaxial cable. Equating the previous two equations, we find that the self-inductance per unit length of the cable is
Equation:

$$\frac{L}{l} = \frac{\mu_0}{2\pi} \ln \frac{R_2}{R_1}.$$

Significance

The inductance per unit length depends only on the inner and outer radii as seen in the result. To increase the inductance, we could either increase the outer radius (R_2) or decrease the inner radius (R_1). In the limit as the two radii become equal, the inductance goes to zero. In this limit, there is no coaxial cable. Also, the magnetic energy per unit length from part (a) is proportional to the square of the current.

Note:
Exercise:

Problem:

Check Your Understanding How much energy is stored in the inductor of [link] after the current reaches its maximum value?

Solution:

0.50 J

Summary

- The energy stored in an inductor U is
  Equation:

$$U=\frac{1}{2}LI^2.$$

- The self-inductance per unit length of coaxial cable is
  Equation:

$$\frac{L}{l}=\frac{\mu_0}{2\pi}\ln\frac{R_2}{R_1}.$$

Conceptual Questions

Exercise:

Problem: Show that $LI^2/2$ has units of energy.

Problems

Exercise:

Problem:

At the instant a current of 0.20 A is flowing through a coil of wire, the energy stored in its magnetic field is $6.0 \times 10^{-3}\,\text{J}$. What is the self-inductance of the coil?

Exercise:

Problem:

Suppose that a rectangular toroid has 2000 windings and a self-inductance of 0.040 H. If $h = 0.10\,\text{m}$, what is the current flowing through a rectangular toroid when the energy in its magnetic field is 2.0×10^{-6} J?

Solution:

0.01 A

Exercise:

Problem:

Solenoid *A* is tightly wound while solenoid *B* has windings that are evenly spaced with a gap equal to the diameter of the wire. The solenoids are otherwise identical. Determine the ratio of the energies stored per unit length of these solenoids when the same current flows through each.

Exercise:

Problem:

A 10-H inductor carries a current of 20 A. How much ice at $0^\circ\ \text{C}$ could be melted by the energy stored in the magnetic field of the inductor? (*Hint*: Use the value $L_\text{f} = 334\ \text{J/g}$ for ice.)

Solution:

6.0 g

Exercise:

Problem:

A coil with a self-inductance of 3.0 H and a resistance of $100\ \Omega$ carries a steady current of 2.0 A. (a) What is the energy stored in the magnetic field of the coil? (b) What is the energy per second dissipated in the resistance of the coil?

Exercise:

Problem:

A current of 1.2 A is flowing in a coaxial cable whose outer radius is five times its inner radius. What is the magnetic field energy stored in a 3.0-m length of the cable?

Solution:

$U_\text{m} = 7.0\ \times\ 10^{-7}\ \text{J}$

Glossary

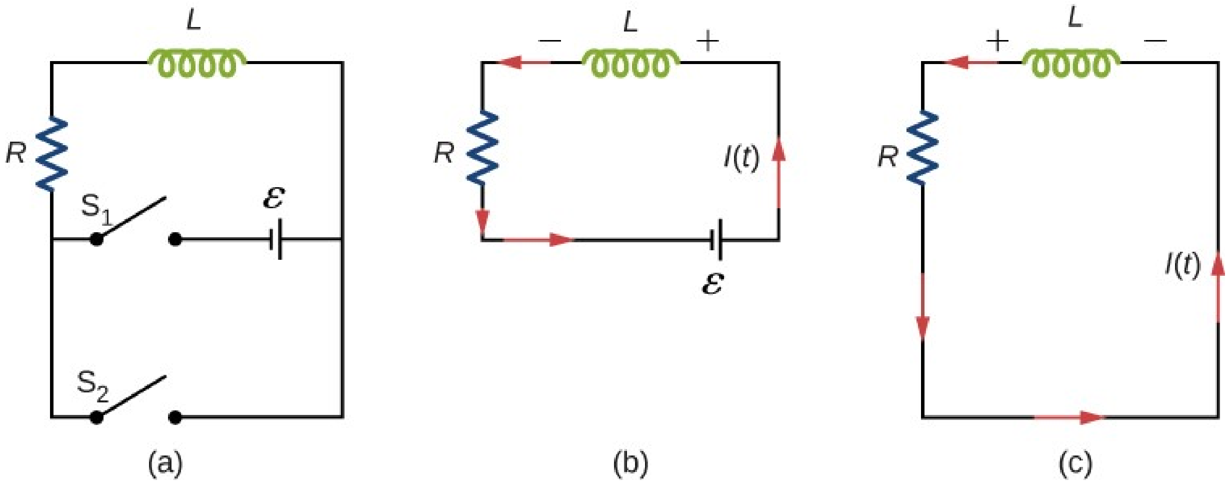
magnetic energy density	energy stored per volume in a magnetic field

RL Circuits

By the end of this section, you will be able to:

- Analyze circuits that have an inductor and resistor in series
- Describe how current and voltage exponentially grow or decay based on the initial conditions

A circuit with resistance and self-inductance is known as an *RL* circuit. [link](a) shows an *RL* circuit consisting of a resistor, an inductor, a constant source of emf, and switches S_1 and S_2. When S_1 is closed, the circuit is equivalent to a single-loop circuit consisting of a resistor and an inductor connected across a source of emf ([link](b)). When S_1 is opened and S_2 is closed, the circuit becomes a single-loop circuit with only a resistor and an inductor ([link](c)).



(a) An *RL* circuit with switches S_1 and S_2. (b) The equivalent circuit with S_1 closed and S_2 open. (c) The equivalent circuit after S_1 is opened and S_2 is closed.

We first consider the *RL* circuit of [link](b). Once S_1 is closed and S_2 is open, the source of emf produces a current in the circuit. If there were no self-inductance in the circuit, the current would rise immediately to a steady value of ε/R. However, from Faraday's law, the increasing current produces an emf $V_L = -L(dI/dt)$ across the inductor. In accordance with Lenz's law, the induced emf counteracts the increase in the current and is directed as shown in the figure. As a result, *I(t)* starts at zero and increases asymptotically to its final value.

Applying Kirchhoff’s loop rule to this circuit, we obtain
Equation:

$$\varepsilon - L\frac{dI}{dt} - IR = 0,$$

which is a first-order differential equation for *I(t)*. Notice its similarity to the equation for a capacitor and resistor in series (See *RC Circuits*). Similarly, the solution to [link] can be found by making substitutions in the equations relating the capacitor to the inductor. This gives

Note:
Equation:

$$I(t) = \frac{\varepsilon}{R}\left(1 - e^{-Rt/L}\right) = \frac{\varepsilon}{R}\left(1 - e^{-t/\tau_L}\right),$$

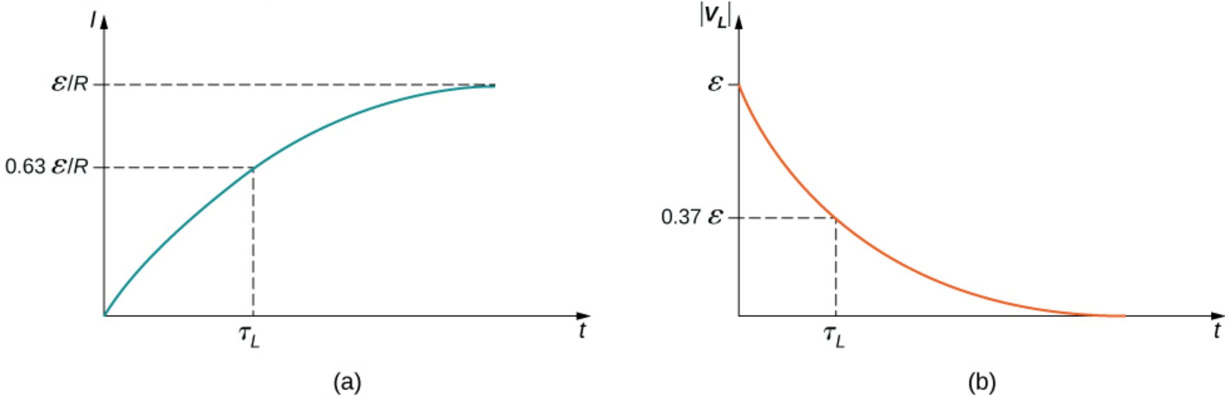
where

Note:
Equation:

$$\tau_L = L/R$$

is the **inductive time constant** of the circuit.

The current *I(t)* is plotted in [link](a). It starts at zero, and as $t \to \infty$, *I(t)* approaches ε / R asymptotically. The induced emf $V_L(t)$ is directly proportional to *dI*/*dt*, or the slope of the curve. Hence, while at its greatest immediately after the switches are thrown, the induced emf decreases to zero with time as the current approaches its final value of ε / R. The circuit then becomes equivalent to a resistor connected across a source of emf.



Time variation of (a) the electric current and (b) the magnitude of the induced voltage across the coil in the circuit of [link](b).

The energy stored in the magnetic field of an inductor is

Equation:

$$U_L = \frac{1}{2} L I^2.$$

Thus, as the current approaches the maximum current ε / R, the stored energy in the inductor increases from zero and asymptotically approaches a maximum of $L(\varepsilon / R)^2 / 2$.

The time constant τ_L tells us how rapidly the current increases to its final value. At $t = \tau_L$, the current in the circuit is, from [link],

Equation:

$$I(\tau_L) = \frac{\varepsilon}{R}(1 - e^{-1}) = 0.63 \frac{\varepsilon}{R},$$

which is 63% of the final value ε/R. The smaller the inductive time constant $\tau_L = L/R$, the more rapidly the current approaches ε/R.

We can find the time dependence of the induced voltage across the inductor in this circuit by using $V_L(t) = -L(dI/dt)$ and [link]:
Equation:

$$V_L(t) = -L\frac{dI}{dt} = -\varepsilon e^{-t/\tau_L}.$$

The magnitude of this function is plotted in [link](b). The greatest value of $L\,(dI/dt)$ is ε; it occurs when *dI/dt* is greatest, which is immediately after S_1 is closed and S_2 is opened. In the approach to steady state, *dI/dt* decreases to zero. As a result, the voltage across the inductor also vanishes as $t \to \infty$.

The time constant τ_L also tells us how quickly the induced voltage decays. At $t = \tau_L$, the magnitude of the induced voltage is
Equation:

$$|V_L(\tau_L)| = \varepsilon e^{-1} = 0.37\varepsilon = 0.37V(0).$$

The voltage across the inductor therefore drops to about 37% of its initial value after one time constant. The shorter the time constant τ_L, the more rapidly the voltage decreases.

After enough time has elapsed so that the current has essentially reached its final value, the positions of the switches in [link](a) are reversed, giving us the circuit in part (c). At $t = 0$, the current in the circuit is $I(0) = \varepsilon/R$. With Kirchhoff’s loop rule, we obtain
Equation:

$$IR + L\frac{dI}{dt} = 0.$$

The solution to this equation is similar to the solution of the equation for a discharging capacitor, with similar substitutions. The current at time *t* is then

Equation:

$$I(t)=\frac{\varepsilon}{R}e^{-t/\tau_L}.$$

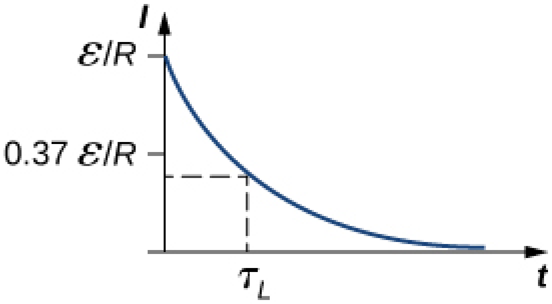
The current starts at $I(0)=\varepsilon/R$ and decreases with time as the energy stored in the inductor is depleted ([link]).

The time dependence of the voltage across the inductor can be determined from $V_L=-L(dI/dt)$:

Equation:

$$V_L(t)=\varepsilon e^{-t/\tau L}.$$

This voltage is initially $V_L(0)=\varepsilon$, and it decays to zero like the current. The energy stored in the magnetic field of the inductor, $LI^2/2$, also decreases exponentially with time, as it is dissipated by Joule heating in the resistance of the circuit.



Time variation of electric current in the *RL* circuit of [link](c). The induced voltage across the coil also decays exponentially.

Example:

An *RL* Circuit with a Source of emf

In the circuit of [link](a), let $\varepsilon = 2.0V$, $R = 4.0\,\Omega,$ and $L = 4.0\,\text{H}.$ With S_1 closed and S_2 open ([link](b)), (a) what is the time constant of the circuit? (b) What are the current in the circuit and the magnitude of the induced emf across the inductor at $t = 0,$ at $t = 2.0\tau_L$, and as $t \to \infty$?

Strategy

The time constant for an inductor and resistor in a series circuit is calculated using [link]. The current through and voltage across the inductor are calculated by the scenarios detailed from [link] and [link].

Solution

a. The inductive time constant is
   Equation:

$$\tau_L = \frac{L}{R} = \frac{4.0\,\text{H}}{4.0\,\Omega} = 1.0\,\text{s}.$$

b. The current in the circuit of [link](b) increases according to [link]:
   Equation:

$$I(t) = \frac{\varepsilon}{R}(1 - e^{-t/\tau_L}).$$

At $t = 0,$
Equation:

$$(1 - e^{-t/\tau_L}) = (1 - 1) = 0;\ \text{so}\ I(0) = 0.$$

At $t = 2.0\tau_L$ and $t \to \infty$, we have, respectively,
Equation:

$$I(2.0\tau_L) = \frac{\varepsilon}{R}(1 - e^{-2.0}) = (0.50\ \text{A})(0.86) = 0.43\ \text{A},$$

and
Equation:

$$I(\infty)=\frac{\varepsilon}{R}=0.50\text{ A}.$$

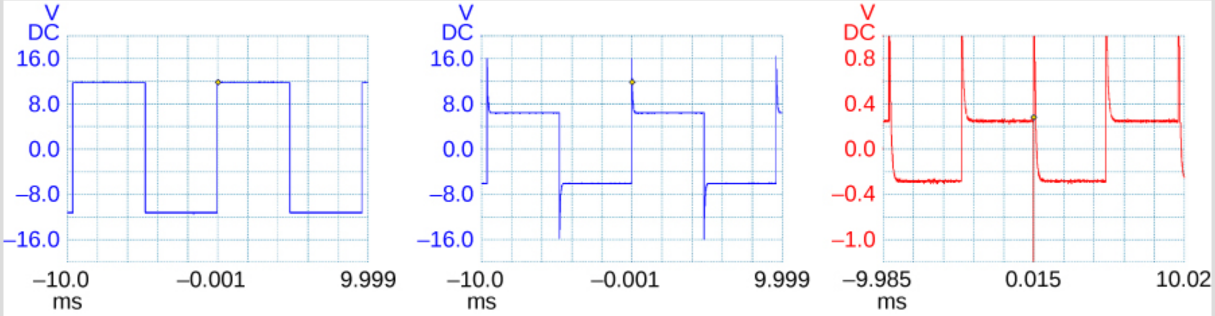
From [link], the magnitude of the induced emf decays as
Equation:

$$|V_L(t)|=\varepsilon e^{-t/\tau_L}.$$

At $t=0, t=2.0\tau_L,$ and as $t\to\infty$, we obtain
Equation:

$$\begin{aligned}|V_L(0)| &= \varepsilon=2.0\text{ V},\\ |V_L(2.0\tau_L)| &= (2.0\text{ V})\,e^{-2.0}=0.27\text{ V}\\ &\text{and}\\ |V_L(\infty)| &= 0.\end{aligned}$$

Significance
If the time of the measurement were much larger than the time constant, we would not see the decay or growth of the voltage across the inductor or resistor. The circuit would quickly reach the asymptotic values for both of these. See [link].



(a) Voltage across the source (b) Voltage across the inductor (c) Voltage across the resistor

A generator in an *RL* circuit produces a square-pulse output in which the voltage oscillates between zero and some set value. These oscilloscope traces show (a) the voltage across the source; (b) the voltage across the inductor; (c) the voltage across the resistor.

Example:

An *RL* Circuit without a Source of emf

After the current in the *RL* circuit of [link] has reached its final value, the positions of the switches are reversed so that the circuit becomes the one shown in [link](c). (a) How long does it take the current to drop to half its initial value? (b) How long does it take before the energy stored in the inductor is reduced to 1.0% of its maximum value?

Strategy

The current in the inductor will now decrease as the resistor dissipates this energy. Therefore, the current falls as an exponential decay. We can also use that same relationship as a substitution for the energy in an inductor formula to find how the energy decreases at different time intervals.

Solution

a. With the switches reversed, the current decreases according to

Equation:

$$I(t)=\frac{\varepsilon}{R}e^{-t/\tau_L}=I(0)e^{-t/\tau_L}.$$

At a time *t* when the current is one-half its initial value, we have

Equation:

$$I(t)=0.50I(0)\text{ so }e^{-t/\tau_L}=0.50,$$

and

Equation:

$$t = -[\ln(0.50)]\tau_L = 0.69(1.0\,\text{s}) = 0.69\,\text{s},$$

where we have used the inductive time constant found in [link].

b. The energy stored in the inductor is given by
Equation:

$$U_L(t) = \frac{1}{2}L[I\,(t)]^2 = \frac{1}{2}L\left(\frac{\varepsilon}{R}e^{-t/\tau_L}\right)^2 = \frac{L\varepsilon^2}{2R^2}e^{-2t/\tau_L}.$$

If the energy drops to 1.0% of its initial value at a time *t*, we have
Equation:

$$U_L(t) = (0.010)U_L(0) \text{ or } \frac{L\varepsilon^2}{2R^2}e^{-2t/\tau_L} = (0.010)\frac{L\varepsilon^2}{2R^2}.$$

Upon canceling terms and taking the natural logarithm of both sides, we obtain
Equation:

$$-\frac{2t}{\tau_L} = \ln(0.010),$$

so
Equation:

$$t = -\frac{1}{2}\tau_L \ln(0.010).$$

Since $\tau_L = 1.0\,\text{s}$, the time it takes for the energy stored in the inductor to decrease to 1.0% of its initial value is
Equation:

$$t = -\frac{1}{2}(1.0\,\text{s})\ln(0.010) = 2.3\,\text{s}.$$

Significance
This calculation only works if the circuit is at maximum current in situation (b) prior to this new situation. Otherwise, we start with a lower initial current, which will decay by the same relationship.

Note:
Exercise:

Problem:

Check Your Understanding Verify that *RC* and *L*/*R* have the dimensions of time.

Note:
Exercise:

Problem:

Check Your Understanding (a) If the current in the circuit of in [link](b) increases to 90% of its final value after 5.0 s, what is the inductive time constant? (b) If $R = 20\,\Omega$, what is the value of the self-inductance? (c) If the 20-Ω resistor is replaced with a 100-Ω resister, what is the time taken for the current to reach 90% of its final value?

Solution:

a. 2.2 s; b. 43 H; c. 1.0 s

Note:
Exercise:

Problem:

Check Your Understanding For the circuit of in [link](b), show that when steady state is reached, the difference in the total energies produced by the battery and dissipated in the resistor is equal to the energy stored in the magnetic field of the coil.

Summary

- When a series connection of a resistor and an inductor—an *RL* circuit—is connected to a voltage source, the time variation of the current is $I(t) = \frac{\varepsilon}{R}(1 - e^{-Rt/L}) = \frac{\varepsilon}{R}(1 - e^{-t/\tau_L})$ (turning on), where the initial current is $I_0 = \varepsilon/R$.
- The characteristic time constant τ is $\tau_L = L/R$, where L is the inductance and R is the resistance.
- In the first time constant τ, the current rises from zero to $0.632I_0$, and to 0.632 of the remainder in every subsequent time interval τ.
- When the inductor is shorted through a resistor, current decreases as $I(t) = \frac{\varepsilon}{R}e^{-t/\tau_L}$ (turning off). Current falls to $0.368I_0$ in the first time interval τ, and to 0.368 of the remainder toward zero in each subsequent time τ.

Conceptual Questions

Exercise:

Problem:

Use Lenz’s law to explain why the initial current in the *RL* circuit of [link](b) is zero.

Solution:

As current flows through the inductor, there is a back current by Lenz's law that is created to keep the net current at zero amps, the initial current.

Exercise:

Problem:

When the current in the *RL* circuit of [link](b) reaches its final value ε / R, what is the voltage across the inductor? Across the resistor?

Exercise:

Problem:

Does the time required for the current in an *RL* circuit to reach any fraction of its steady-state value depend on the emf of the battery?

Solution:

no

Exercise:

Problem:

An inductor is connected across the terminals of a battery. Does the current that eventually flows through the inductor depend on the internal resistance of the battery? Does the time required for the current to reach its final value depend on this resistance?

Exercise:

Problem:

At what time is the voltage across the inductor of the *RL* circuit of [link](b) a maximum?

Solution:

At $t = 0$, or when the switch is first thrown.

Exercise:

Problem:

In the simple *RL* circuit of [link](b), can the emf induced across the inductor ever be greater than the emf of the battery used to produce the current?

Exercise:

Problem:

If the emf of the battery of [link](b) is reduced by a factor of 2, by how much does the steady-state energy stored in the magnetic field of the inductor change?

Solution:

1/4

Exercise:

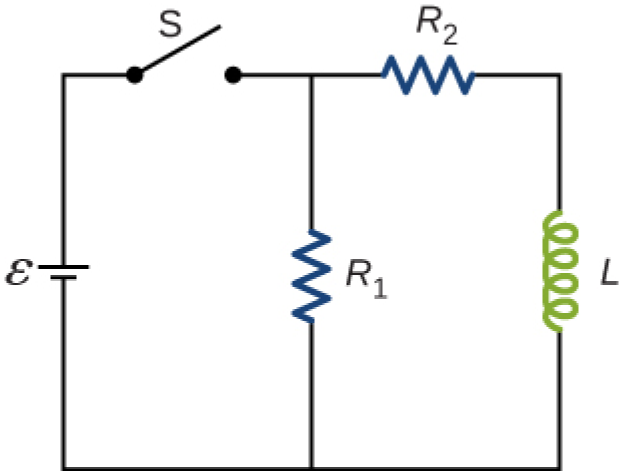
Problem:

A steady current flows through a circuit with a large inductive time constant. When a switch in the circuit is opened, a large spark occurs across the terminals of the switch. Explain.

Exercise:

Problem:

Describe how the currents through R_1 and R_2 shown below vary with time after switch S is closed.



Solution:

Initially, $I_{R1}=\frac{\varepsilon}{R_1}$ and $I_{R2}=0$, and after a long time has passed, $I_{R1}=\frac{\varepsilon}{R_1}$ and $I_{R2}=\frac{\varepsilon}{R_2}$.

Exercise:

Problem: Discuss possible practical applications of *RL* circuits.

Problems

Exercise:

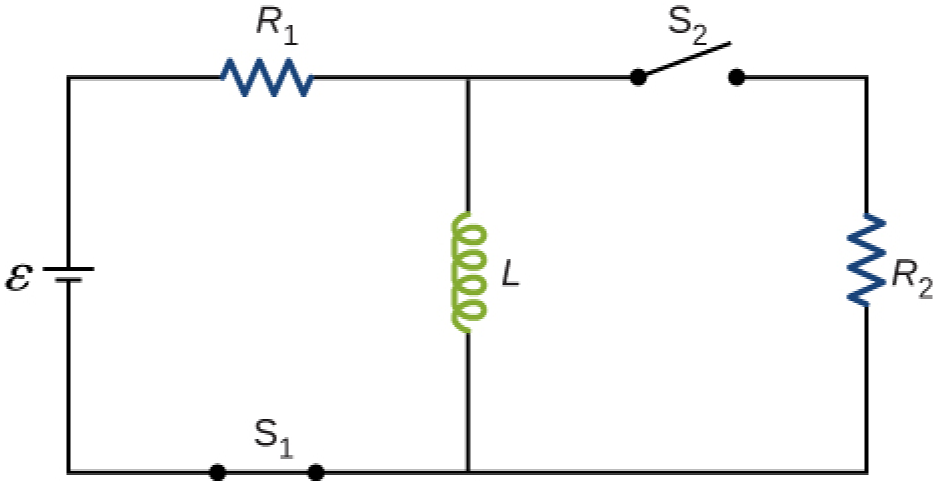
Problem:

In [link], $\varepsilon=12\text{ V}$, $L=20\text{ mH}$, and $R=5.0\,\Omega$. Determine (a) the time constant of the circuit, (b) the initial current through the resistor, (c) the final current through the resistor, (d) the current through the resistor when $t=2\tau_L$, and (e) the voltages across the inductor and the resistor when $t=2\tau_L$.

Exercise:

Problem:

For the circuit shown below, $\varepsilon = 20\ \text{V}$, $L = 4.0\ \text{mH}$, and $R = 5.0\ \Omega$. After steady state is reached with S_1 closed and S_2 open, S_2 is closed and immediately thereafter $(\text{at } t = 0)$ S_1 is opened. Determine (a) the current through L at $t = 0$, (b) the current through L at $t = 4.0 \times 10^{-4}$ s, and (c) the voltages across L and R_1 at $t = 4.0 \times 10^{-4}$ s. $R_1 = R_2 = R$.



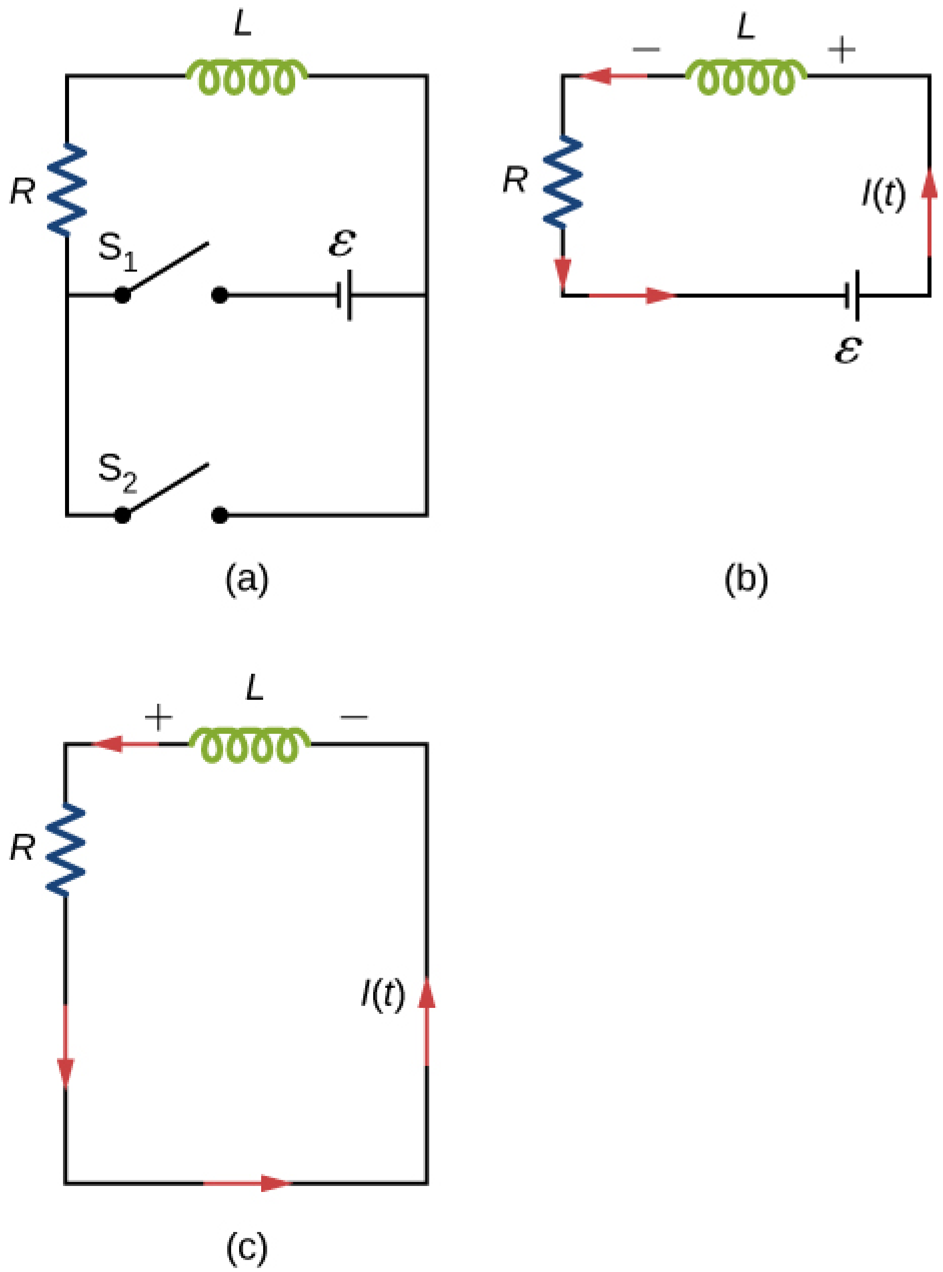
Solution:

a. 4.0 A; b. 2.4 A; c. on R: $V = 12\ \text{V}$; on L: $V = 7.9\ \text{V}$

Exercise:

Problem:

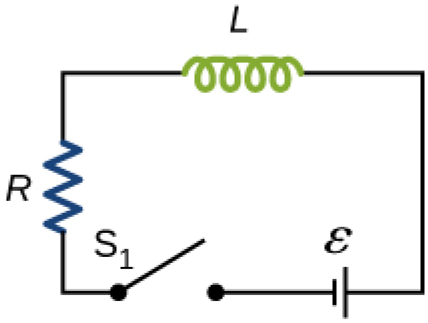
The current in the *RL* circuit shown here increases to 40% of its steady-state value in 2.0 s. What is the time constant of the circuit?


L
R
S1
ℰ
S2
(a)
−
L
+
R
I(t)
ℰ
(b)
+
L
−
R
I(t)
(c)

Exercise:

Problem:

How long after switch S_1 is thrown does it take the current in the circuit shown to reach half its maximum value? Express your answer in terms of the time constant of the circuit.


L
R
S_1
$\mathcal{E}$

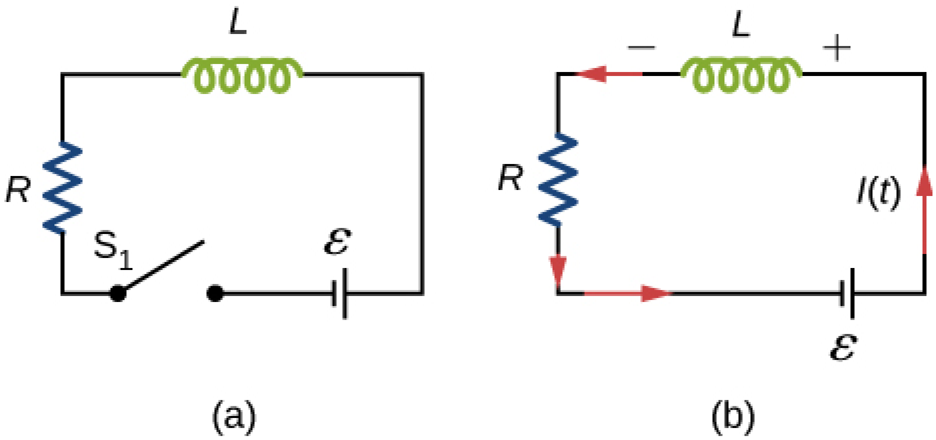
Solution:

0.69τ

Exercise:

Problem:

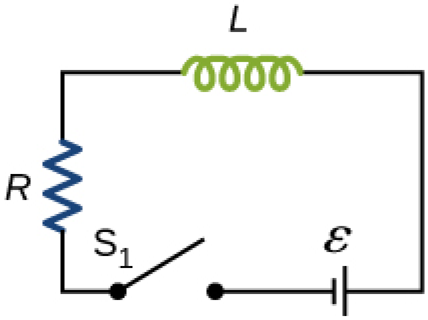
Examine the circuit shown below in part (a). Determine *dI/dt* at the instant after the switch is thrown in the circuit of (a), thereby producing the circuit of (b). Show that if *I* were to continue to increase at this initial rate, it would reach its maximum ε/R in one time constant.


L
R
S_1
$\mathcal{E}$
(a)
–
L
+
R
I(t)
$\mathcal{E}$
(b)

Exercise:

Problem:

The current in the *RL* circuit shown below reaches half its maximum value in 1.75 ms after the switch S_1 is thrown. Determine (a) the time constant of the circuit and (b) the resistance of the circuit if $L = 250\,\mathrm{mH}$.



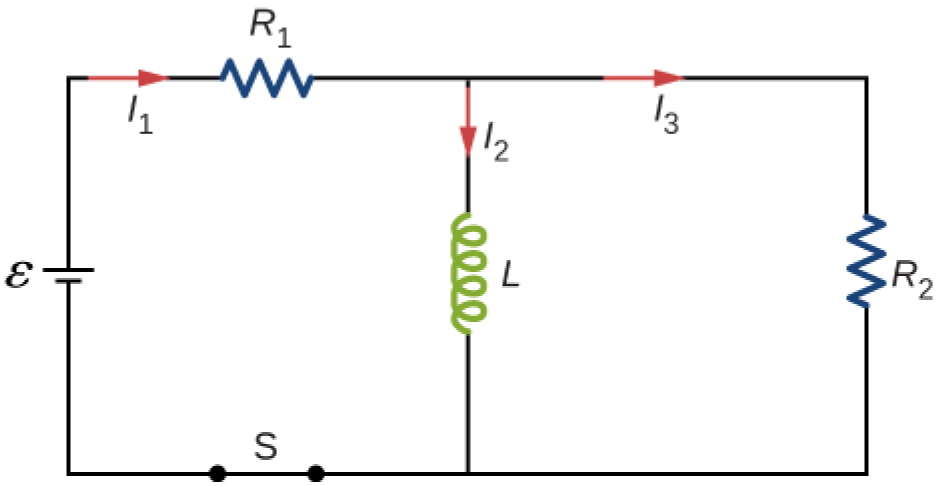
Solution:

a. 2.52 ms; b. 99.2 Ω

Exercise:

Problem:

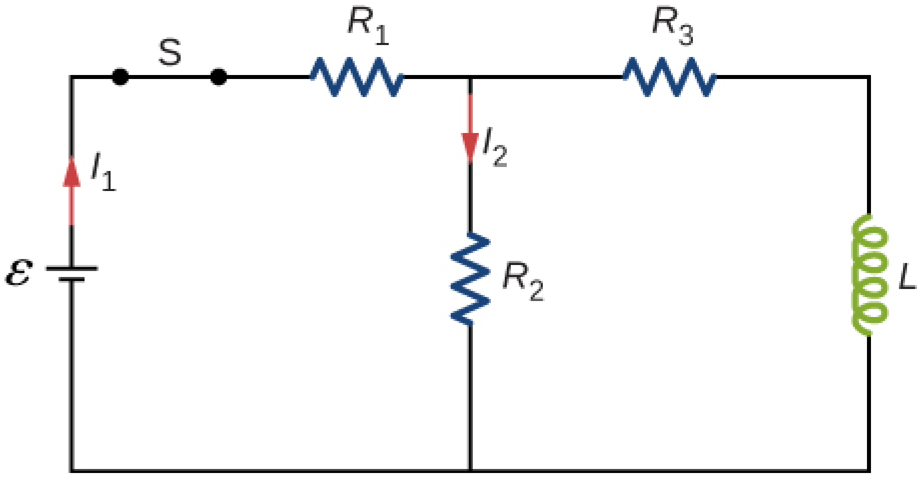
Consider the circuit shown below. Find $I_1, I_2,$ and I_3 when (a) the switch S is first closed, (b) after the currents have reached steady-state values, and (c) at the instant the switch is reopened (after being closed for a long time).


R_1
I_1
I_2
I_3
ε
L
R_2
S

Exercise:

Problem:

For the circuit shown below, $\varepsilon = 50\ \text{V}$, $R_1 = 10\ \Omega$,, $R_2 = R_3 = \ 19.4\ \Omega$,, and $L = 2.0\ \text{mH}$. Find the values of I_1 and I_2 (a) immediately after switch S is closed, (b) a long time after S is closed, (c) immediately after S is reopened, and (d) a long time after S is reopened.


S
R_1
R_3
I_1
I_2
ε
R_2
L

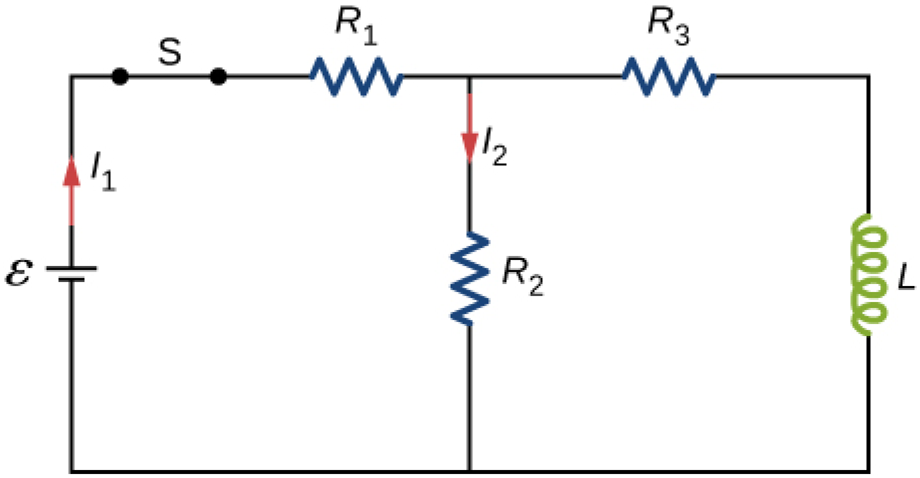
Solution:

a. $I_1 = I_2 = 1.7\ A$; b. $I_1 = 2.73\ A, I_2 = 1.36\ A$; c. $I_1 = 0, I_2 = 0.54\ A$; d. $I_1 = I_2 = 0$

Exercise:

Problem:

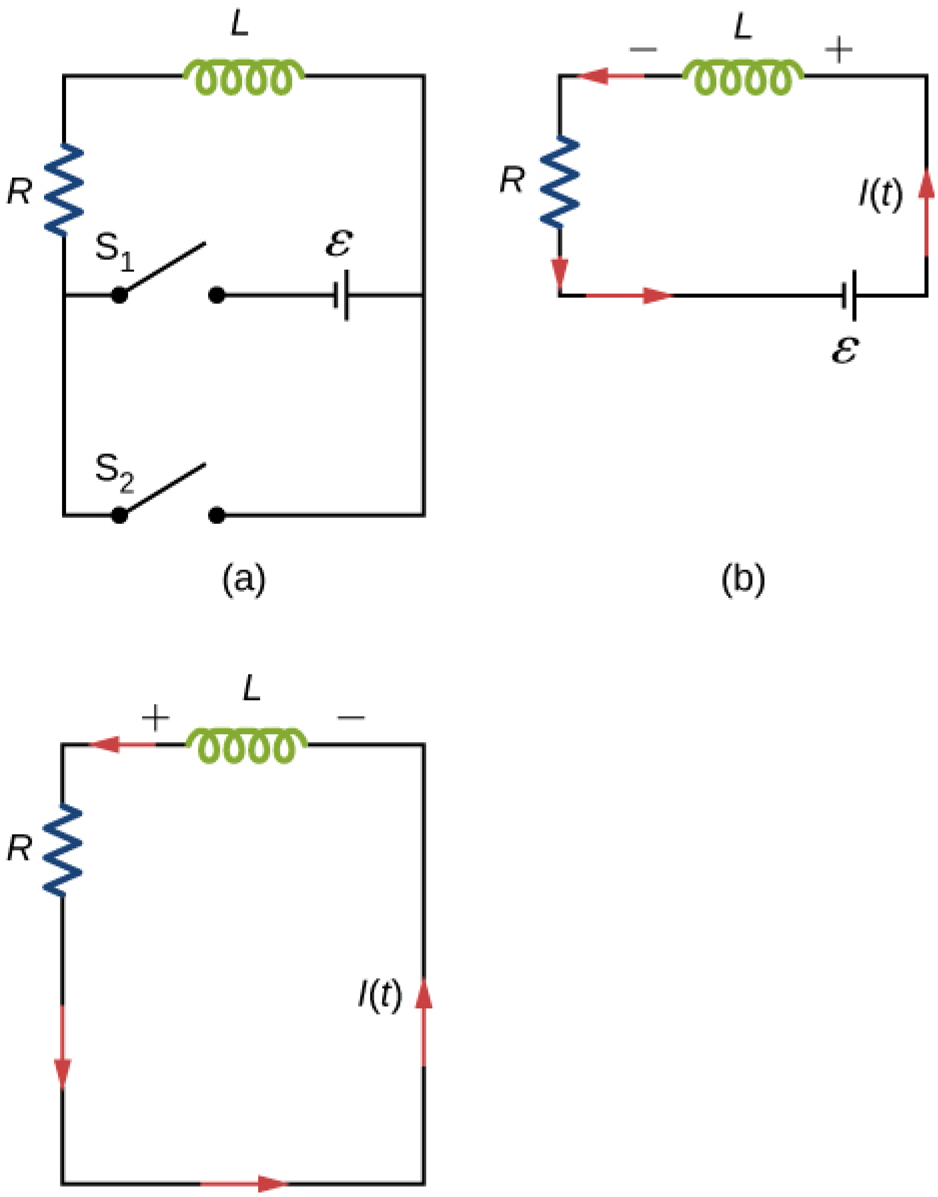
For the circuit shown below, find the current through the inductor 2.0×10^{-5} s after the switch is reopened.



Exercise:

Problem:

Show that for the circuit shown below, the initial energy stored in the inductor, $LI^2(0)/2$, is equal to the total energy eventually dissipated in the resistor, $\int_0^\infty I^2(t)R\,dt$.



(a) (b)

(c)

Solution:

proof

Glossary

inductive time constant

denoted by τ, the characteristic time given by quantity L/R of a particular series RL circuit

Oscillations in an LC Circuit

By the end of this section, you will be able to:

- Explain why charge or current oscillates between a capacitor and inductor, respectively, when wired in series
- Describe the relationship between the charge and current oscillating between a capacitor and inductor wired in series

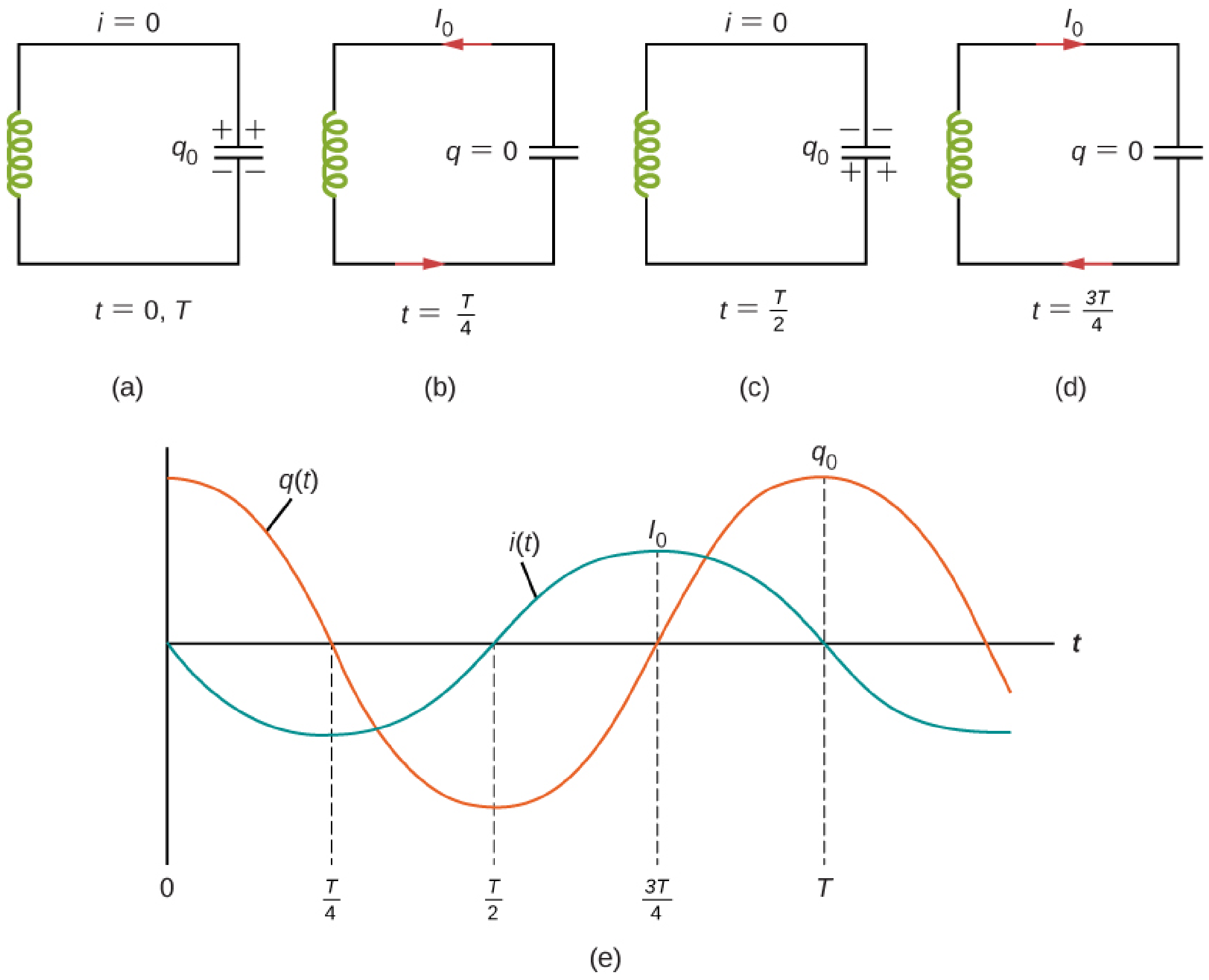
It is worth noting that both capacitors and inductors store energy, in their electric and magnetic fields, respectively. A circuit containing both an inductor (L) and a capacitor (C) can oscillate without a source of emf by shifting the energy stored in the circuit between the electric and magnetic fields. Thus, the concepts we develop in this section are directly applicable to the exchange of energy between the electric and magnetic fields in electromagnetic waves, or light. We start with an idealized circuit of zero resistance that contains an inductor and a capacitor, an ***LC* circuit**.

An *LC* circuit is shown in [link]. If the capacitor contains a charge q_0 before the switch is closed, then all the energy of the circuit is initially stored in the electric field of the capacitor ([link](a)). This energy is

Equation:

$$U_C = \frac{1}{2}\frac{q_0^2}{C}.$$

When the switch is closed, the capacitor begins to discharge, producing a current in the circuit. The current, in turn, creates a magnetic field in the inductor. The net effect of this process is a transfer of energy from the capacitor, with its diminishing electric field, to the inductor, with its increasing magnetic field.



$i = 0$
q_0
$t = 0, T$
(a)
I_0
$q = 0$
$t = \frac{T}{4}$
(b)
$i = 0$
q_0
$t = \frac{T}{2}$
(c)
I_0
$q = 0$
$t = \frac{3T}{4}$
(d)
$q(t)$
$i(t)$
I_0
q_0
t
0
$\frac{T}{4}$
$\frac{T}{2}$
$\frac{3T}{4}$
T
(e)

(a–d) The oscillation of charge storage with changing directions of current in an LC circuit. (e) The graphs show the distribution of charge and current between the capacitor and inductor.

In [link](b), the capacitor is completely discharged and all the energy is stored in the magnetic field of the inductor. At this instant, the current is at its maximum value I_0 and the energy in the inductor is

Equation:

$$U_L = \frac{1}{2} L I_0^2.$$

Since there is no resistance in the circuit, no energy is lost through Joule heating; thus, the maximum energy stored in the capacitor is equal to the maximum energy stored at a later time in the inductor:
Equation:

$$\frac{1}{2}\frac{q_0^2}{C}=\frac{1}{2}LI_0^2.$$

At an arbitrary time when the capacitor charge is *q(t)* and the current is *i(t)*, the total energy *U* in the circuit is given by
Equation:

$$\frac{q^2(t)}{2C}+\frac{Li^2(t)}{2}.$$

Because there is no energy dissipation,
Equation:

$$U=\frac{1}{2}\frac{q^2}{C}+\frac{1}{2}Li^2=\frac{1}{2}\frac{q_0^2}{C}=\frac{1}{2}LI_0^2.$$

After reaching its maximum I_0, the current *i(t)* continues to transport charge between the capacitor plates, thereby recharging the capacitor. Since the inductor resists a change in current, current continues to flow, even though the capacitor is discharged. This continued current causes the capacitor to charge with opposite polarity. The electric field of the capacitor increases while the magnetic field of the inductor diminishes, and the overall effect is a transfer of energy from the inductor *back* to the capacitor. From the law of energy conservation, the maximum charge that the capacitor re-acquires is q_0. However, as [link](c) shows, the capacitor plates are charged *opposite* to what they were initially.

When fully charged, the capacitor once again transfers its energy to the inductor until it is again completely discharged, as shown in [link](d). Then, in the last part of this cyclic process, energy flows back to the capacitor, and the initial state of the circuit is restored.

We have followed the circuit through one complete cycle. Its electromagnetic oscillations are analogous to the mechanical oscillations of a mass at the end of a spring. In this latter case, energy is transferred back and forth between the mass, which has kinetic energy $mv^2/2$, and the spring, which has potential energy $kx^2/2$. With the absence of friction in the mass-spring system, the oscillations would continue indefinitely. Similarly, the oscillations of an *LC* circuit with no resistance would continue forever if undisturbed; however, this ideal zero-resistance *LC* circuit is not practical, and any *LC* circuit will have at least a small resistance, which will radiate and lose energy over time.

The frequency of the oscillations in a resistance-free *LC* circuit may be found by analogy with the mass-spring system. For the circuit, $i(t) = dq(t)/dt$, the total electromagnetic energy *U* is

Equation:

$$U = \frac{1}{2}Li^2 + \frac{1}{2}\frac{q^2}{C}.$$

For the mass-spring system, $v(t) = dx(t)/dt$, the total mechanical energy *E* is

Equation:

$$E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2.$$

The equivalence of the two systems is clear. To go from the mechanical to the electromagnetic system, we simply replace *m* by *L*, *v* by *i*, *k* by 1/*C*, and *x* by *q*. Now *x(t)* is given by

Equation:

$$x(t) = A\cos(\omega t + \phi)$$

where $\omega = \sqrt{k/m}$. Hence, the charge on the capacitor in an *LC* circuit is given by

Note:

Equation:

$$q\left(t\right)={q}_{0}\cos(\omega t+\phi)$$

where the angular frequency of the oscillations in the circuit is

Note:
Equation:

$$\omega =\sqrt{\frac{1}{LC}}.$$

Finally, the current in the *LC* circuit is found by taking the time derivative of *q(t)*:

Note:
Equation:

$$i(t)=\frac{dq(t)}{dt}=-\omega {q}_{0}\sin(\omega t+\phi).$$

The time variations of q and I are shown in [link](e) for $\phi =0$.

Example:
An *LC* Circuit

In an *LC* circuit, the self-inductance is 2.0×10^{-2} H and the capacitance is 8.0×10^{-6} F. At $t = 0$, all of the energy is stored in the capacitor, which has charge 1.2×10^{-5} C. (a) What is the angular frequency of the oscillations in the circuit? (b) What is the maximum current flowing through circuit? (c) How long does it take the capacitor to become completely discharged? (d) Find an equation that represents *q(t)*.

Strategy

The angular frequency of the *LC* circuit is given by [link]. To find the maximum current, the maximum energy in the capacitor is set equal to the maximum energy in the inductor. The time for the capacitor to become discharged if it is initially charged is a quarter of the period of the cycle, so if we calculate the period of the oscillation, we can find out what a quarter of that is to find this time. Lastly, knowing the initial charge and angular frequency, we can set up a cosine equation to find $q(t)$.

Solution

a. From [link], the angular frequency of the oscillations is

   Equation:

$$\omega = \sqrt{\frac{1}{LC}} = \sqrt{\frac{1}{(2.0 \times 10^{-2}\,\mathrm{H})(8.0 \times 10^{-6}\,\mathrm{F})}} = 2.5 \times 10^{3}\,\mathrm{rad/s}.$$

b. The current is at its maximum I_0 when all the energy is stored in the inductor. From the law of energy conservation,

   Equation:

$$\frac{1}{2}LI_0^2 = \frac{1}{2}\frac{q_0^2}{C},$$

   so

   Equation:

$$I_0 = \sqrt{\frac{1}{LC}}q_0 = (2.5 \times 10^{3}\,\mathrm{rad/s})(1.2 \times 10^{-5}\,\mathrm{C}) = 3.0 \times 10^{-2}\,\mathrm{A}.$$

   This result can also be found by an analogy to simple harmonic motion, where current and charge are the velocity and position of an oscillator.

c. The capacitor becomes completely discharged in one-fourth of a cycle, or during a time *T*/4, where *T* is the period of the oscillations. Since
Equation:

$$T=\frac{2\pi}{\omega}=\frac{2\pi}{2.5\ \times\ 10^{3}\,\text{rad/s}}=2.5\ \times\ 10^{-3}\,\text{s},$$

the time taken for the capacitor to become fully discharged is $(2.5\ \times\ 10^{-3}\,\text{s})/4=6.3\ \times\ 10^{-4}\,\text{s}.$

d. The capacitor is completely charged at $t=0$, so $q(0)=q_0$. Using [link], we obtain
Equation:

$$q(0)=q_0=q_0\cos\phi.$$

Thus, $\phi=0$, and
Equation:

$$q(t)=(1.2\ \times\ 10^{-5}\,\text{C})\cos(2.5\ \times\ 10^{3}t).$$

Significance
The energy relationship set up in part (b) is not the only way we can equate energies. At most times, some energy is stored in the capacitor and some energy is stored in the inductor. We can put both terms on each side of the equation. By examining the circuit only when there is no charge on the capacitor or no current in the inductor, we simplify the energy equation.

Note:
Exercise:

Problem:

Check Your Understanding The angular frequency of the oscillations in an LC circuit is 2.0×10^3 rad/s. (a) If $L = 0.10\,\text{H}$, what is C? (b) Suppose that at $t = 0$, all the energy is stored in the inductor. What is the value of ϕ? (c) A second identical capacitor is connected in parallel with the original capacitor. What is the angular frequency of this circuit?

Solution:

a. $2.5\mu\text{F}$; b. $\pi/2\,\text{rad}$ or $3\pi/2\,\text{rad}$; c. $1.4 \times 10^3\,\text{rad/s}$

Summary

- The energy transferred in an oscillatory manner between the capacitor and inductor in an LC circuit occurs at an angular frequency $\omega = \sqrt{\frac{1}{LC}}$.
- The charge and current in the circuit are given by **Equation:**

$$\begin{aligned} q(t) &= q_0 \cos(\omega t + \phi), \\ i(t) &= -\omega q_0 \sin(\omega t + \phi). \end{aligned}$$

Conceptual Questions

Exercise:

Problem:

Do Kirchhoff’s rules apply to circuits that contain inductors and capacitors?

Solution:

yes

Exercise:

Problem: Can a circuit element have both capacitance and inductance?

Exercise:

Problem:

In an *LC* circuit, what determines the frequency and the amplitude of the energy oscillations in either the inductor or capacitor?

Solution:

The amplitude of energy oscillations depend on the initial energy of the system. The frequency in a *LC* circuit depends on the values of inductance and capacitance.

Problems

Exercise:

Problem:

A 5000-pF capacitor is charged to 100 V and then quickly connected to an 80-mH inductor. Determine (a) the maximum energy stored in the magnetic field of the inductor, (b) the peak value of the current, and (c) the frequency of oscillation of the circuit.

Exercise:

Problem:

The self-inductance and capacitance of an *LC* circuit are 0.20 mH and 5.0 pF. What is the angular frequency at which the circuit oscillates?

Solution:

$\omega = 3.2 \times 10^7 \text{ rad/s}$

Exercise:

Problem:

What is the self-inductance of an *LC* circuit that oscillates at 60 Hz when the capacitance is 10 μF?

Exercise:

Problem:

In an oscillating *LC* circuit, the maximum charge on the capacitor is $2.0\ \times\ 10^{-6}$ C and the maximum current through the inductor is 8.0 mA. (a) What is the period of the oscillations? (b) How much time elapses between an instant when the capacitor is uncharged and the next instant when it is fully charged?

Solution:

a. $1.57\ \times\ 10^{-6}$ s; b. $3.93\ \times\ 10^{-7}$ s

Exercise:

Problem:

The self-inductance and capacitance of an oscillating *LC* circuit are $L = 20\ \text{mH}$ and $C = 1.0\ \mu\text{F}$, respectively. (a) What is the frequency of the oscillations? (b) If the maximum potential difference between the plates of the capacitor is 50 V, what is the maximum current in the circuit?

Exercise:

Problem:

In an oscillating *LC* circuit, the maximum charge on the capacitor is q_m. Determine the charge on the capacitor and the current through the inductor when energy is shared equally between the electric and magnetic fields. Express your answer in terms of q_m, *L*, and *C*.

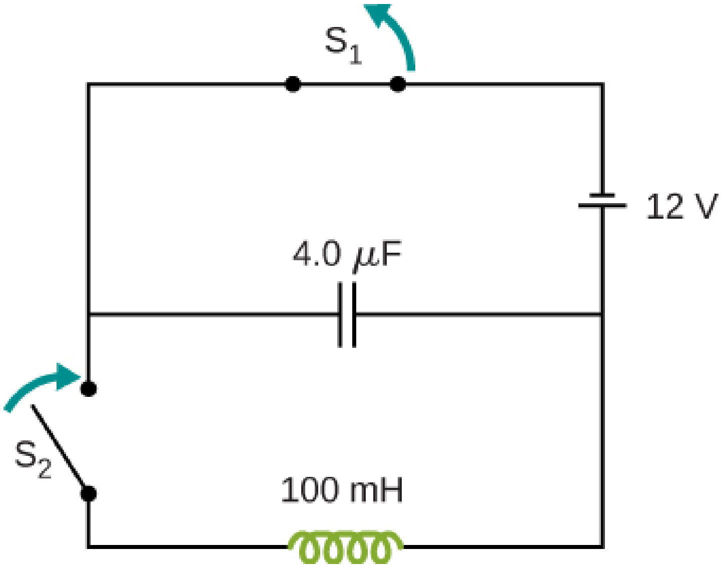
Solution:

$q = \frac{q_m}{\sqrt{2}}, I = \frac{q_m}{\sqrt{2LC}}$

Exercise:

Problem:

In the circuit shown below, S_1 is opened and S_2 is closed simultaneously. Determine (a) the frequency of the resulting oscillations, (b) the maximum charge on the capacitor, (c) the maximum current through the inductor, and (d) the electromagnetic energy of the oscillating circuit.



Exercise:

Problem:

An *LC* circuit in an AM tuner (in a car stereo) uses a coil with an inductance of 2.5 mH and a variable capacitor. If the natural frequency of the circuit is to be adjustable over the range 540 to 1600 kHz (the AM broadcast band), what range of capacitance is required?

Solution:

$$
\begin{aligned}
C &= \frac{1}{4\pi^2 f^2 L} & & \\
f_1 &= 540\ \text{Hz}; & C_1 &= 3.5 \times 10^{-11}\ \text{F} \\
f_2 &= 1600\ \text{Hz}; & C_2 &= 4.0 \times 10^{-12}\ \text{F}
\end{aligned}
$$

Glossary

LC circuit

circuit composed of an ac source, inductor, and capacitor

RLC Series Circuits

By the end of this section, you will be able to:

- Determine the angular frequency of oscillation for a resistor, inductor, capacitor (RLC) series circuit
- Relate the RLC circuit to a damped spring oscillation

When the switch is closed in the ***RLC* circuit** of [link](a), the capacitor begins to discharge and electromagnetic energy is dissipated by the resistor at a rate i^2R. With U given by [link], we have

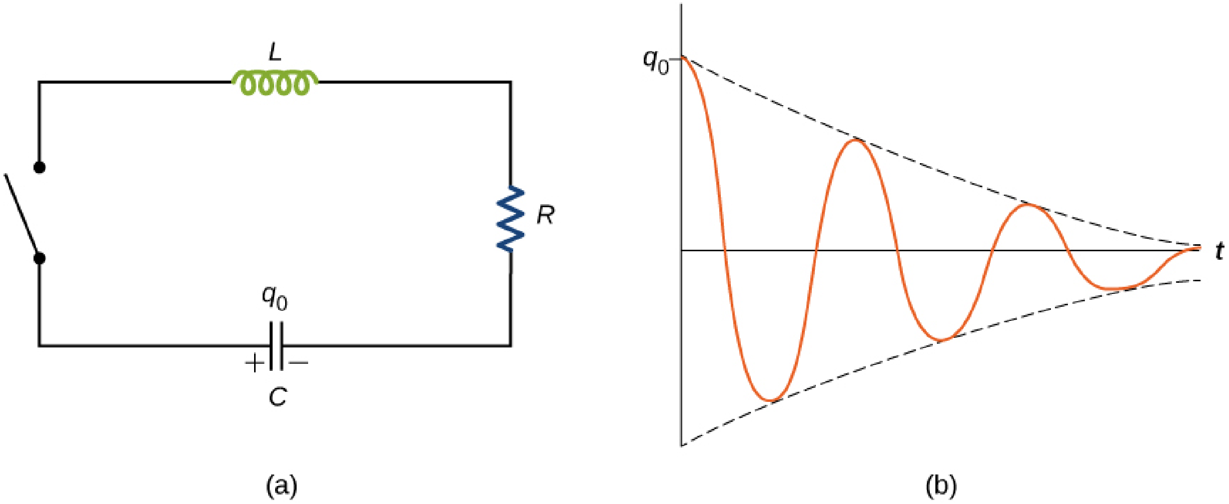
Equation:

$$\frac{dU}{dt} = \frac{q}{C}\frac{dq}{dt} + Li\frac{di}{dt} = -i^2R$$

where i and q are time-dependent functions. This reduces to

Equation:

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = 0.$$



(a) An *RLC* circuit. Electromagnetic oscillations begin when the switch is closed. The capacitor is fully charged initially. (b) Damped oscillations of the capacitor charge are shown in this curve of charge versus time, or q versus t. The capacitor contains a charge q_0 before the switch is closed.

This equation is analogous to

Equation:

$$m\frac{d^2x}{dt^2}+b\frac{dx}{dt}+kx=0,$$

which is the equation of motion for a *damped mass-spring system* (you first encountered this equation in Oscillations). As we saw in that chapter, it can be shown that the solution to this differential equation takes three forms, depending on whether the angular frequency of the undamped spring is greater than, equal to, or less than $b/2m$. Therefore, the result can be underdamped ($\sqrt{k/m}>b/2m$), critically damped ($\sqrt{k/m}=b/2m$), or overdamped ($\sqrt{k/m}<b/2m$). By analogy, the solution $q(t)$ to the *RLC* differential equation has the same feature. Here we look only at the case of under-damping. By replacing m by L, b by R, k by $1/C$, and x by q in [link], and assuming $\sqrt{1/LC}>R/2L$, we obtain

Note:

Equation:

$$q(t)=q_0e^{-Rt/2L}\cos(\omega' t+\phi)$$

where the angular frequency of the oscillations is given by

Note:

Equation:

$$\omega' = \sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}$$

This underdamped solution is shown in [link](b). Notice that the amplitude of the oscillations decreases as energy is dissipated in the resistor. [link] can be confirmed experimentally by measuring the voltage across the capacitor as a function of time. This voltage, multiplied by the capacitance of the capacitor, then gives $q(t)$.

Note:
Try an interactive circuit construction kit that allows you to graph current and voltage as a function of time. You can add inductors and capacitors to work with any combination of R, L, and C circuits with both dc and ac sources.

Note:
Try out a circuit-based java applet website that has many problems with both dc and ac sources that will help you practice circuit problems.

Note:
Exercise:

Problem:

Check Your Understanding In an *RLC* circuit, $L = 5.0\,\text{mH}, C = 6.0\mu\text{F},\ \text{and}\ R = 200\,\Omega.$ (a) Is the circuit underdamped, critically damped, or overdamped? (b) If the circuit starts oscillating with a charge of $3.0\ \times\ 10^{-3}\,\text{C}$ on the capacitor, how much energy has been dissipated in the resistor by the time the oscillations cease?

Solution:

a. overdamped; b. 0.75 J

Summary

- The underdamped solution for the capacitor charge in an *RLC* circuit is
  Equation:

$$q(t) = q_0 e^{-Rt/2L} \cos(\omega\prime t + \phi).$$

- The angular frequency given in the underdamped solution for the *RLC* circuit is
  Equation:

$$\omega' = \sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2}.$$

Key Equations

Mutual inductance by flux	$M=\frac{N_2\Phi_{21}}{I_1}=\frac{N_1\Phi_{12}}{I_2}$
Mutual inductance in circuits	$\varepsilon_1=-M\frac{dI_2}{dt}$
Self-inductance in terms of magnetic flux	$N\Phi_{\mathrm{m}}=LI$
Self-inductance in terms of emf	$\varepsilon=-L\frac{dI}{dt}$
Self-inductance of a solenoid	$L_{\text{solenoid}}=\frac{\mu_0 N^2 A}{l}$
Self-inductance of a toroid	$L_{\text{toroid}}=\frac{\mu_0 N^2 h}{2\pi}\ln\frac{R_2}{R_1}.$
Energy stored in an inductor	$U=\frac{1}{2}LI^2$
Current as a function of time for a *RL* circuit	$I(t)=\frac{\varepsilon}{R}\left(1-e^{-t/\tau_L}\right)$
Time constant for a *RL* circuit	$\tau_L=L/R$
Charge oscillation in *LC* circuits	$q\,(t)=q_0\cos(\omega t+\phi)$
Angular frequency in *LC* circuits	$\omega=\sqrt{\frac{1}{LC}}$
Current oscillations in *LC* circuits	$i(t)=-\omega q_0\sin(\omega t+\phi)$
Charge as a function of time in *RLC* circuit	$q(t)=q_0e^{-Rt/2L}\cos(\omega' t+\phi)$
Angular frequency in *RLC* circuit	$\omega'=\sqrt{\frac{1}{LC}-\left(\frac{R}{2L}\right)^2}$

Conceptual Questions

Exercise:

Problem:

When a wire is connected between the two ends of a solenoid, the resulting circuit can oscillate like an *RLC* circuit. Describe what causes the capacitance in this circuit.

Exercise:

Problem:

Describe what effect the resistance of the connecting wires has on an oscillating *LC* circuit.

Solution:

This creates an *RLC* circuit that dissipates energy, causing oscillations to decrease in amplitude slowly or quickly depending on the value of resistance.

Exercise:

Problem:

Suppose you wanted to design an *LC* circuit with a frequency of 0.01 Hz. What problems might you encounter?

Exercise:

Problem:

A radio receiver uses an *RLC* circuit to pick out particular frequencies to listen to in your house or car without hearing other unwanted frequencies. How would someone design such a circuit?

Solution:

You would have to pick out a resistance that is small enough so that only one station at a time is picked up, but big enough so that the tuner doesn’t have to be set at exactly the correct frequency. The inductance or capacitance would have to be varied to tune into the station however practically speaking, variable capacitors are a lot easier to build in a circuit.

Problems

Exercise:

Problem:

In an oscillating *RLC* circuit, $R = 5.0\,\Omega, L = 5.0\,\text{mH}, \text{ and } C = 500\,\mu\text{F}$. What is the angular frequency of the oscillations?

Exercise:

Problem:

In an oscillating *RLC* circuit with $L = 10\,\text{mH}, C = 1.5\,\mu\text{F}, \text{ and } R = 2.0\,\Omega$, how much time elapses before the amplitude of the oscillations drops to half its initial value?

Solution:

6.9 ms

Exercise:

Problem:

What resistance *R* must be connected in series with a 200-mH inductor of the resulting *RLC* oscillating circuit is to decay to 50% of its initial value of charge in 50 cycles? To 0.10% of its initial value in 50 cycles?

Additional Problems

Exercise:

Problem:

Show that the self-inductance per unit length of an infinite, straight, thin wire is infinite.

Solution:

Let *a* equal the radius of the long, thin wire, *r* the location where the magnetic field is measured, and *R* the upper limit of the problem where we will take *R* as it approaches infinity.

$$\text{Outside,}\quad B = \frac{\mu_0 I}{2\pi r}\ \text{Inside,}\quad B = \frac{\mu_0 I r}{2\pi a^2}$$

proof

$$U = \frac{\mu_0 I^2 l}{4\pi}\left(\frac{1}{4} + \ln\frac{R}{a}\right)$$

$$\text{So,}\quad \frac{2U}{I^2} = \frac{\mu_0 l}{2\pi}\left(\frac{1}{4} + \ln\frac{R}{a}\right)\ \text{and}\quad L = \infty$$

Exercise:

Problem:

Two long, parallel wires carry equal currents in opposite directions. The radius of each wire is *a*, and the distance between the centers of the wires is *d*. Show that if the magnetic flux within the wires themselves can be ignored, the self-inductance of a length *l* of such a pair of wires is

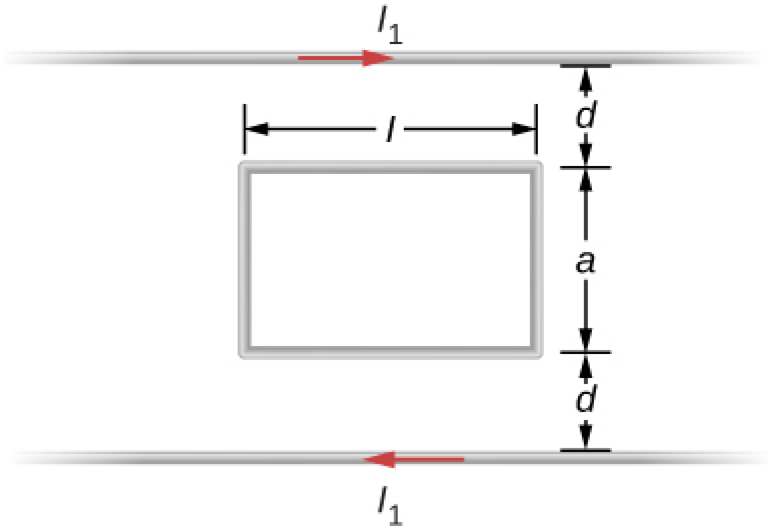
$$L = \frac{\mu_0 l}{\pi}\ln\frac{d-a}{a}.$$

(*Hint*: Calculate the magnetic flux through a rectangle of length *l* between the wires and then use $L = N\Phi/I$.)

Exercise:

Problem:

A small, rectangular single loop of wire with dimensions *l*, and *a* is placed, as shown below, in the plane of a much larger, rectangular single loop of wire. The two short sides of the larger loop are so far from the smaller loop that their magnetic fields over the smaller fields over the smaller loop can be ignored. What is the mutual inductance of the two loops?



Solution:

$$M=\frac{\mu_0 l}{\pi}\ln\frac{d+a}{d}$$

Exercise:

Problem:

Suppose that a cylindrical solenoid is wrapped around a core of iron whose magnetic susceptibility is ***x***. Using [link], show that the self-inductance of the solenoid is given by

$$L=\frac{(1+x)\mu_0 N^2 A}{l},$$

where *l* is its length, *A* its cross-sectional area, and *N* its total number of turns.

Exercise:

Problem:

A solenoid with 4×10^7 turns/m has an iron core placed in it whose magnetic susceptibility is 4.0×10^3. (a) If a current of 2.0 A flows through the solenoid, what is the magnetic field in the iron core? (b) What is the effective surface current formed by the aligned atomic current loops in the iron core? (c) What is the self-inductance of the filled solenoid?

Solution:

a. 100 T; b. 2 A; c. 0.50 H

Exercise:

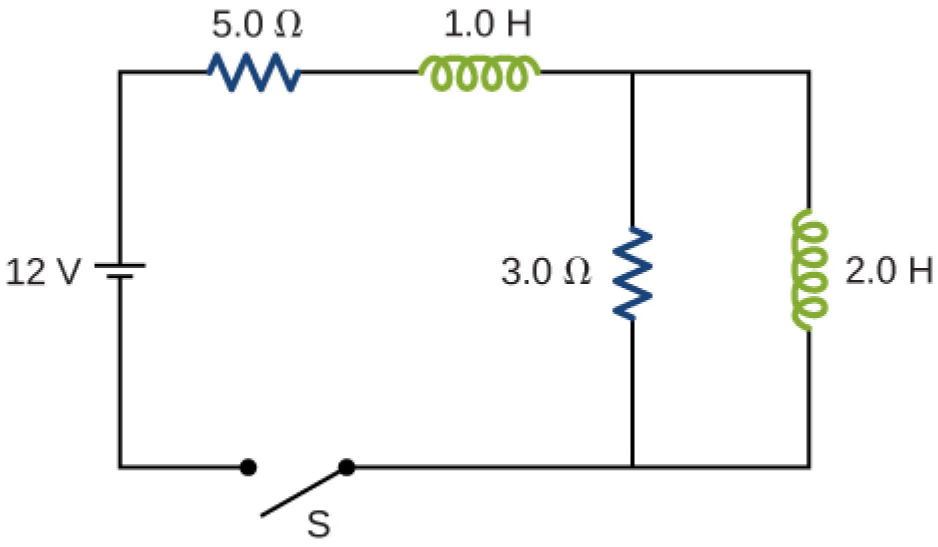
Problem:

A rectangular toroid with inner radius $R_1 = 7.0\,\text{cm}$, outer radius $R_2 = 9.0\,\text{cm}$, height $h = 3.0$, and $N = 3000$ turns is filled with an iron core of magnetic susceptibility 5.2×10^3. (a) What is the self-inductance of the toroid? (b) If the current through the toroid is 2.0 A, what is the magnetic field at the center of the core? (c) For this same 2.0-A current, what is the effective surface current formed by the aligned atomic current loops in the iron core?

Exercise:

Problem:

The switch S of the circuit shown below is closed at $t = 0$. Determine (a) the initial current through the battery and (b) the steady-state current through the battery.



Solution:

a. 0 A; b. 2.4 A

Exercise:

Problem:

In an oscillating *RLC* circuit, $R = 7.0\,\Omega, L = 10\,\text{mH},$ and $C = 3.0\,\mu\text{F}$. Initially, the capacitor has a charge of $8.0\,\mu\text{C}$ and the current is zero. Calculate the charge on the capacitor (a) five cycles later and (b) 50 cycles later.

Exercise:

Problem:

A 25.0-H inductor has 100 A of current turned off in 1.00 ms. (a) What voltage is induced to oppose this? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

a. 2.50×10^6 V; (b) The voltage is so extremely high that arcing would occur and the current would not be reduced so rapidly. (c) It is not reasonable to shut off such a large current in such a large inductor in such an extremely short time.

Challenge Problems

Exercise:

Problem:

A coaxial cable has an inner conductor of radius a, and outer thin cylindrical shell of radius b. A current I flows in the inner conductor and returns in the outer conductor. The self-inductance of the structure will depend on how the current in the inner cylinder tends to be distributed. Investigate the following two extreme cases. (a) Let current in the inner conductor be distributed only on the surface and find the self-inductance. (b) Let current in the inner cylinder be distributed uniformly over its cross-section and find the self-inductance. Compare with your results in (a).

Exercise:

Problem:

In a damped oscillating circuit the energy is dissipated in the resistor. The Q-factor is a measure of the persistence of the oscillator against the dissipative loss. (a) Prove that for a lightly damped circuit the energy, U, in the circuit decreases according to the following equation.

$\frac{dU}{dt} = -2\beta U$, where $\beta = \frac{R}{2L}$.

(b) Using the definition of the Q-factor as energy divided by the loss over the next cycle, prove that Q-factor of a lightly damped oscillator as defined in this problem is

$$Q \equiv \frac{U_{\text{begin}}}{\Delta U_{\text{one cycle}}} = \frac{1}{2\pi R}\sqrt{\frac{L}{C}}.$$

(*Hint:* For (b), to obtain Q, divide E at the beginning of one cycle by the change ΔE over the next cycle.)

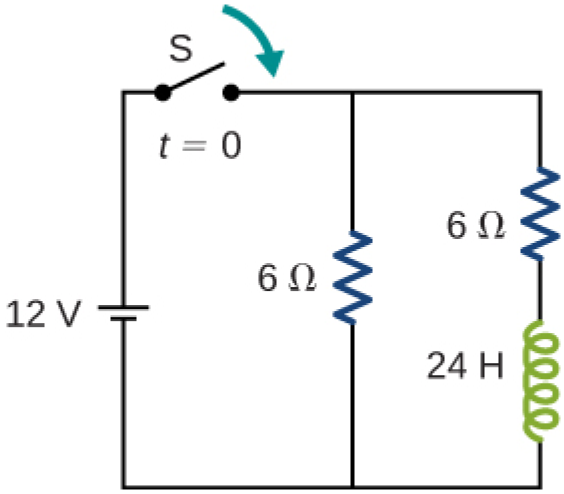
Solution:

proof

Exercise:

Problem:

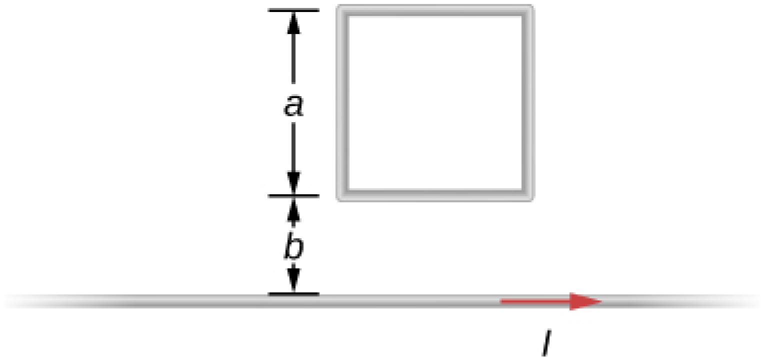
The switch in the circuit shown below is closed at $t = 0\,\text{s}$. Find currents through (a) R_1, (b) R_2, and (c) the battery as function of time.



Exercise:

Problem:

A square loop of side 2 cm is placed 1 cm from a long wire carrying a current that varies with time at a constant rate of 3 A/s as shown below. (a) Use Ampère’s law and find the magnetic field. (b) Determine the magnetic flux through the loop. (c) If the loop has a resistance of $3\,\Omega$, how much induced current flows in the loop?



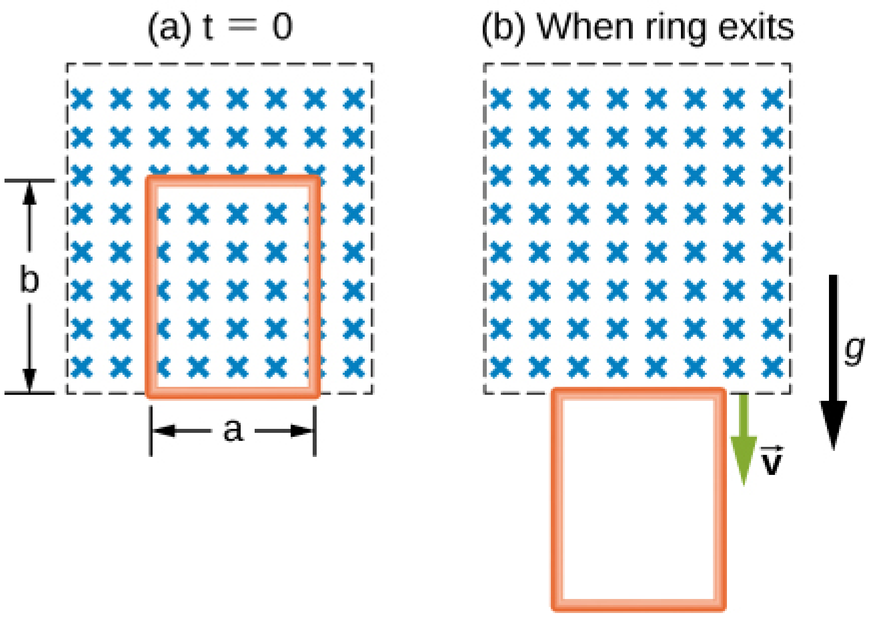
Solution:

a. $\frac{dB}{dt} = 6 \times 10^{-6}\,\text{T/s}$; b. $\Phi = \frac{\mu_0 a I}{2\pi} \ln \left(\frac{a+b}{b}\right)$; c. 4.4 nA

Exercise:

Problem:

A rectangular copper ring, of mass 100 g and resistance $0.2\,\Omega$, is in a region of uniform magnetic field that is perpendicular to the area enclosed by the ring and horizontal to Earth’s surface. The ring is let go from rest when it is at the edge of the nonzero magnetic field region (see below). (a) Find its speed when the ring just exits the region of uniform magnetic field. (b) If it was let go at $t = 0$, what is the time when it exits the region of magnetic field for the following values: $a = 25\,\text{cm},\ b = 50\,\text{cm},\ B = 3\,\text{T},$ and $g = 9.8\,\text{m/s}^2$? Assume the magnetic field of the induced current is negligible compared to 3 T.


(a) t = 0
b
a
(b) When ring exits
g
$\vec{v}$

Glossary

RLC circuit
circuit with an ac source, resistor, inductor, and capacitor all in series.

Introduction class="introduction"

The current we draw into our houses is an alternating current (ac). Power lines transmit ac to our neighborhoods, where local power stations and transformers distribute it to our homes. In this chapter, we discuss how a transformer works and how it allows us to transmit power at very high voltages and minimal heating losses across the lines.

Electric power is delivered to our homes by alternating current (ac) through high-voltage transmission lines. As explained in Transformers, transformers can then change the amplitude of the alternating potential difference to a more useful form. This lets us transmit power at very high voltages, minimizing resistive heating losses in the lines, and then furnish that power to homes at lower, safer voltages. Because constant potential differences are unaffected by transformers, this capability is more difficult to achieve with direct-current transmission.

In this chapter, we use Kirchhoff’s laws to analyze four simple circuits in which ac flows. We have discussed the use of the resistor, capacitor, and inductor in circuits with batteries. These components are also part of ac circuits. However, because ac is required, the constant source of emf supplied by a battery is replaced by an ac voltage source, which produces an oscillating emf.

AC Sources

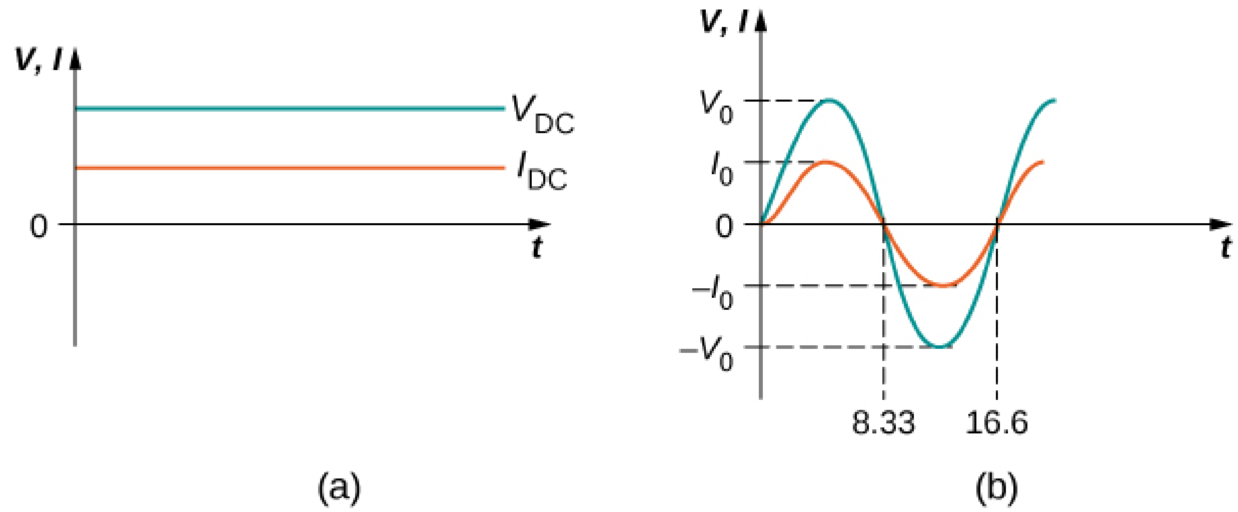
By the end of the section, you will be able to:

- Explain the differences between direct current (dc) and alternating current (ac)
- Define characteristic features of alternating current and voltage, such as the amplitude or peak and the frequency

Most examples dealt with so far in this book, particularly those using batteries, have constant-voltage sources. Thus, once the current is established, it is constant. **Direct current (dc)** is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit.

Most well-known applications, however, use a time-varying voltage source. **Alternating current (ac)** is the flow of electric charge that periodically reverses direction. An ac is produced by an alternating emf, which is generated in a power plant, as described in Induced Electric Fields. If the ac source varies periodically, particularly sinusoidally, the circuit is known as an ac circuit. Examples include the commercial and residential power that serves so many of our needs.

The ac voltages and frequencies commonly used in businesses and homes vary around the world. In a typical house, the potential difference between the two sides of an electrical outlet alternates sinusoidally with a frequency of 60 or 50 Hz and an amplitude of 170 or 311 V, depending on whether you live in the United States or Europe, respectively. Most people know the potential difference for electrical outlets is 120 V or 220 V in the US or Europe, but as explained later in the chapter, these voltages are not the peak values given here but rather are related to the common voltages we see in our electrical outlets. [link] shows graphs of voltage and current versus time for typical dc and ac power in the United States.



(a) The dc voltage and current are constant in time, once the current is established. (b) The voltage and current versus time are quite different for ac power. In this example, which shows 60-Hz ac power and time *t* in milliseconds, voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of ac sources differ greatly.

Suppose we hook up a resistor to an ac voltage source and determine how the voltage and current vary in time across the resistor. [link] shows a schematic of a simple circuit with an ac voltage source. The voltage fluctuates sinusoidally with time at a fixed frequency, as shown, on either the battery terminals or the resistor. Therefore, the **ac voltage**, or the “voltage at a plug,” can be given by

Note:
Equation:

$$v = V_0 \sin \omega t,$$

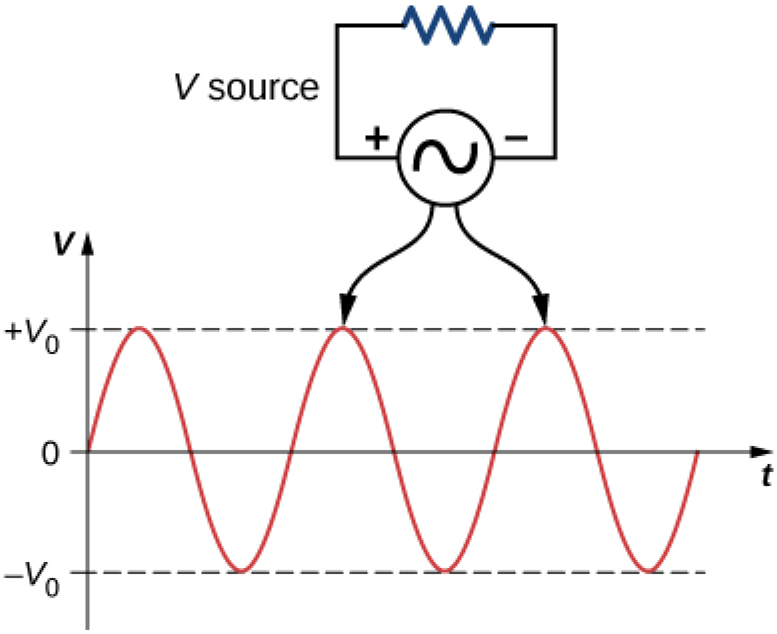
where *v* is the voltage at time *t*, V_0 is the peak voltage, and ω is the angular frequency in radians per second. For a typical house in the United States, $V_0 = 170 \text{ V}$ and $\omega = 120\pi \text{ rad/s}$, whereas in Europe, $V_0 = 311 \text{ V}$ and $\omega = 100\pi \text{ rad/s}$.

For this simple resistance circuit, $I = V/R$, so the **ac current**, meaning the current that fluctuates sinusoidally with time at a fixed frequency, is

Note:
Equation:

$$i = I_0 \sin \omega t,$$

where *i* is the current at time *t* and I_0 is the peak current and is equal to V_0/R. For this example, the voltage and current are said to be in phase, meaning that their sinusoidal functional forms have peaks, troughs, and nodes in the same place. They oscillate in sync with each other, as shown in [link](b). In these equations, and throughout this chapter, we use lowercase letters (such as *i*) to indicate instantaneous values and capital letters (such as *I*) to indicate maximum, or peak, values.



The potential difference *V* between the terminals of an ac voltage source fluctuates, so the source and the resistor have ac sine waves on top of each other. The mathematical expression for *v* is given by $v = V_0 \sin \omega t$.

Current in the resistor alternates back and forth just like the driving voltage, since $I = V/R$. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see the stroboscopic effect of ac.

Note:
Exercise:

Problem:

Check Your Understanding If a European ac voltage source is considered, what is the time difference between the zero crossings on an ac voltage-versus-time graph?

Solution:

10 ms

Summary

- Direct current (dc) refers to systems in which the source voltage is constant.
- Alternating current (ac) refers to systems in which the source voltage varies periodically, particularly sinusoidally.
- The voltage source of an ac system puts out a voltage that is calculated from the time, the peak voltage, and the angular frequency.
- In a simple circuit, the current is found by dividing the voltage by the resistance. An ac current is calculated using the peak current (determined by dividing the peak voltage by the resistance), the angular frequency, and the time.

Conceptual Questions

Exercise:

Problem:

What is the relationship between frequency and angular frequency?

Solution:

Angular frequency is 2π times frequency.

Problems

Exercise:

Problem:

Write an expression for the output voltage of an ac source that has an amplitude of 12 V and a frequency of 200 Hz.

Glossary

ac current
current that fluctuates sinusoidally with time at a fixed frequency

ac voltage
voltage that fluctuates sinusoidally with time at a fixed frequency

alternating current (ac)
flow of electric charge that periodically reverses direction

direct current (dc)
flow of electric charge in only one direction

Simple AC Circuits

By the end of the section, you will be able to:

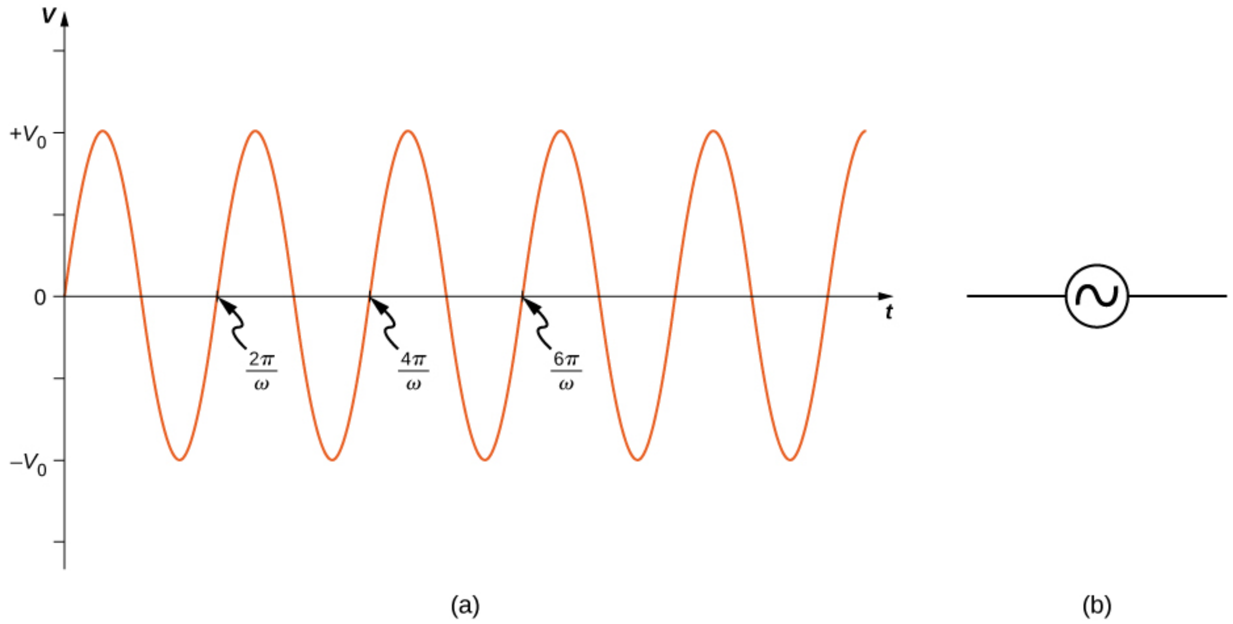
- Interpret phasor diagrams and apply them to ac circuits with resistors, capacitors, and inductors
- Define the reactance for a resistor, capacitor, and inductor to help understand how current in the circuit behaves compared to each of these devices

In this section, we study simple models of ac voltage sources connected to three circuit components: (1) a resistor, (2) a capacitor, and (3) an inductor. The power furnished by an ac voltage source has an emf given by

Equation:

$$v(t) = V_0 \sin \omega t,$$

as shown in [link]. This sine function assumes we start recording the voltage when it is $v = 0\,\text{V}$ at a time of $t = 0\,\text{s}$. A phase constant may be involved that shifts the function when we start measuring voltages, similar to the phase constant in the waves we studied in Waves. However, because we are free to choose when we start examining the voltage, we can ignore this phase constant for now. We can measure this voltage across the circuit components using one of two methods: (1) a quantitative approach based on our knowledge of circuits, or (2) a graphical approach that is explained in the coming sections.



(a) The output $v(t) = V_0 \sin \omega t$ of an ac generator. (b) Symbol used to represent an ac voltage source in a circuit diagram.

Resistor

First, consider a resistor connected across an ac voltage source. From Kirchhoff's loop rule, the instantaneous voltage across the resistor of [link] (a) is

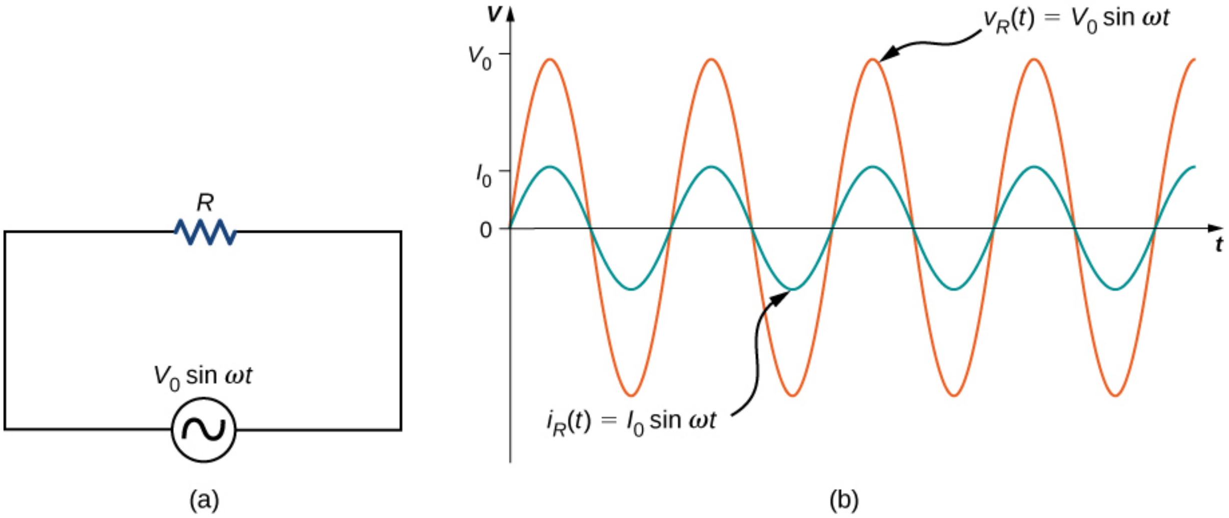
Equation:

$$v_R(t) = V_0 \sin \omega t$$

and the instantaneous current through the resistor is

Equation:

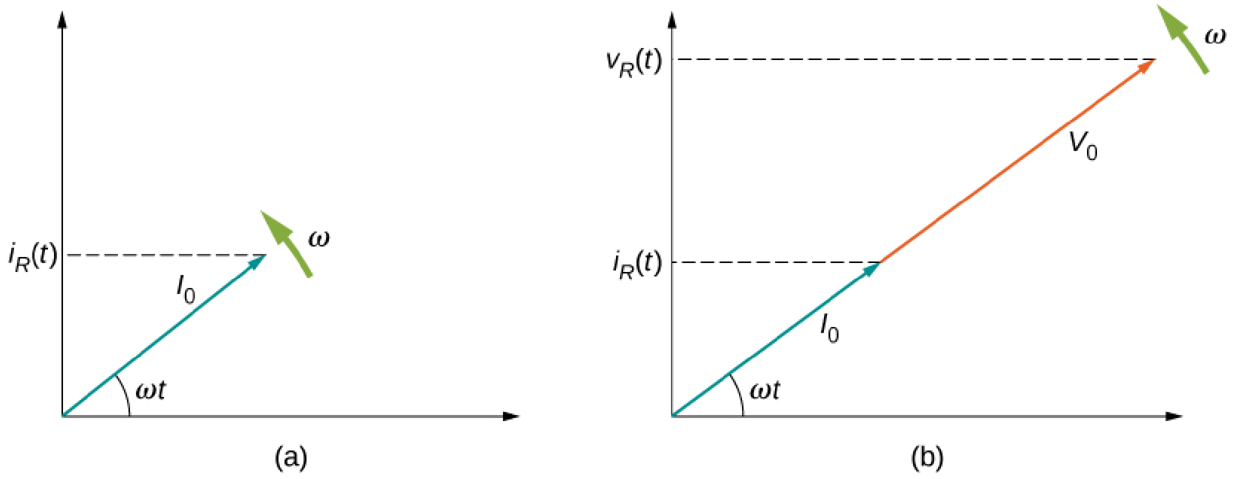
$$i_R(t) = \frac{v_R(t)}{R} = \frac{V_0}{R} \sin \omega t = I_0 \sin \omega t.$$



(a) A resistor connected across an ac voltage source. (b) The current $i_R(t)$ through the resistor and the voltage $v_R(t)$ across the resistor. The two quantities are in phase.

Here, $I_0 = V_0/R$ is the amplitude of the time-varying current. Plots of $i_R(t)$ and $v_R(t)$ are shown in [link](b). Both curves reach their maxima and minima at the same times, that is, the current through and the voltage across the resistor are in phase.

Graphical representations of the phase relationships between current and voltage are often useful in the analysis of ac circuits. Such representations are called *phasor diagrams*. The phasor diagram for $i_R(t)$ is shown in [link](a), with the current on the vertical axis. The arrow (or phasor) is rotating counterclockwise at a constant angular frequency ω, so we are viewing it at one instant in time. If the length of the arrow corresponds to the current amplitude I_0, the projection of the rotating arrow onto the vertical axis is $i_R(t) = I_0 \sin \omega t$, which is the instantaneous current.



(a) The phasor diagram representing the current through the resistor of [link]. (b) The phasor diagram representing both $i_R(t)$ and $v_R(t)$.

The vertical axis on a phasor diagram could be either the voltage or the current, depending on the phasor that is being examined. In addition, several quantities can be depicted on the same phasor diagram. For example, both the current $i_R(t)$ and the voltage $v_R(t)$ are shown in the diagram of [link](b). Since they have the same frequency and are in phase, their phasors point in the same direction and rotate together. The relative lengths of the two phasors are arbitrary because they represent different quantities; however, the ratio of the lengths of the two phasors can be represented by the resistance, since one is a voltage phasor and the other is a current phasor.

Capacitor

Now let’s consider a capacitor connected across an ac voltage source. From Kirchhoff’s loop rule, the instantaneous voltage across the capacitor of [link](a) is

Equation:

$$v_C(t) = V_0 \sin \omega t.$$

Recall that the charge in a capacitor is given by $Q = CV$. This is true at any time measured in the ac cycle of voltage. Consequently, the instantaneous charge on the capacitor is

Equation:

$$q(t) = Cv_C(t) = CV_0 \sin \omega t.$$

Since the current in the circuit is the rate at which charge enters (or leaves) the capacitor,

Equation:

$$i_C(t) = \frac{dq(t)}{dt} = \omega CV_0 \cos \omega t = I_0 \cos \omega t,$$

where $I_0 = \omega CV_0$ is the current amplitude. Using the trigonometric relationship $\cos \omega t = \sin (\omega t + \pi/2)$, we may express the instantaneous current as

Equation:

$$i_C(t) = I_0 \sin \left(\omega t + \frac{\pi}{2}\right).$$

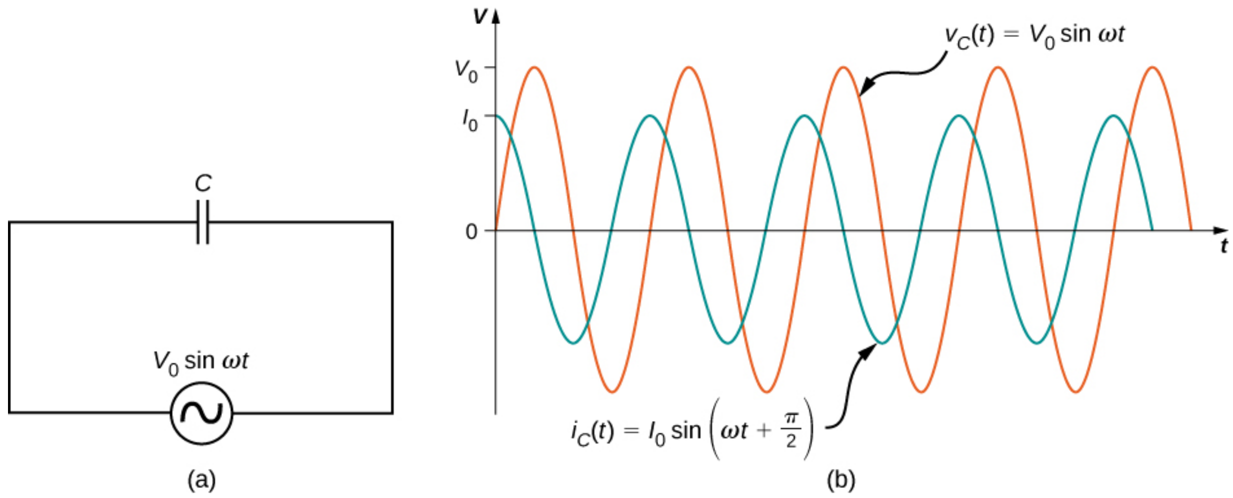
Dividing V_0 by I_0, we obtain an equation that looks similar to Ohm’s law:

Note:

Equation:

$$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C.$$

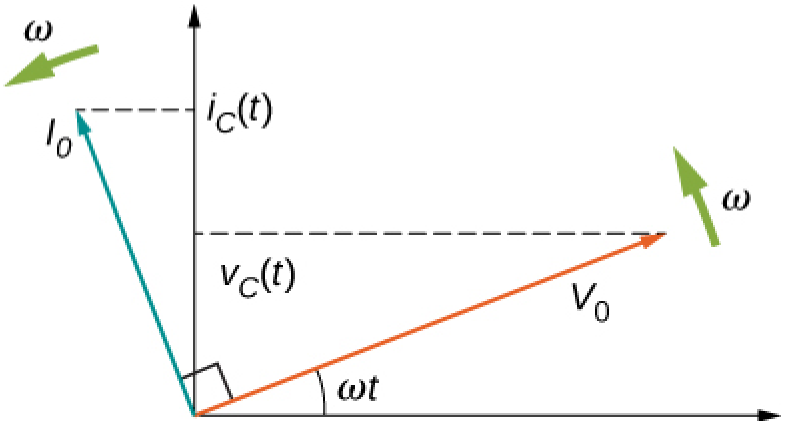
The quantity X_C is analogous to resistance in a dc circuit in the sense that both quantities are a ratio of a voltage to a current. As a result, they have the same unit, the ohm. Keep in mind, however, that a capacitor stores and discharges electric energy, whereas a resistor dissipates it. The quantity X_C is known as the **capacitive reactance** of the capacitor, or the opposition of a capacitor to a change in current. It depends inversely on the frequency of the ac source—high frequency leads to low capacitive reactance.


C
$V_0 \sin \omega t$
(a)
v
V_0
I_0
0
t
$v_C(t) = V_0 \sin \omega t$
$i_C(t) = I_0 \sin\left(\omega t + \frac{\pi}{2}\right)$
(b)

(a) A capacitor connected across an ac generator. (b) The current $i_C(t)$ through the capacitor and the voltage $v_C(t)$ across the capacitor. Notice that $i_C(t)$ leads $v_C(t)$ by $\pi/2$ rad.

A comparison of the expressions for $v_C(t)$ and $i_C(t)$ shows that there is a phase difference of $\pi/2\,\text{rad}$ between them. When these two quantities are plotted together, the current peaks a quarter cycle (or $\pi/2\,\text{rad}$) ahead of the voltage, as illustrated in [link](b). The current through a capacitor leads the voltage across a capacitor by $\pi/2\,\text{rad}$, or a quarter of a cycle.

The corresponding phasor diagram is shown in [link]. Here, the relationship between $i_C(t)$ and $v_C(t)$ is represented by having their phasors rotate at the same angular frequency, with the current phasor leading by $\pi/2\,\text{rad}$.



The phasor diagram for the capacitor of [link]. The current phasor leads the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

To this point, we have exclusively been using peak values of the current or voltage in our discussion, namely, I_0 and V_0. However, if we average out the values of current or voltage, these values are zero. Therefore, we often use a second convention called the root mean square value, or rms value, in discussions of current and voltage. The rms operates in reverse of the terminology. First, you square the function, next, you take the mean, and then, you find the square root. As a result, the rms values of current and voltage are not zero. Appliances and devices are commonly quoted with rms values for their operations, rather than peak values. We indicate rms values with a subscript attached to a capital letter (such as I_{rms}).

Although a capacitor is basically an open circuit, an **rms current**, or the root mean square of the current, appears in a circuit with an ac voltage applied to a capacitor. Consider that

Note:
Equation:

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}},$$

where I_0 is the peak current in an ac system. The **rms voltage**, or the root mean square of the voltage, is

Note:
Equation:

$$V_{\text{rms}} = \frac{V_0}{\sqrt{2}},$$

where V_0 is the peak voltage in an ac system. The rms current appears because the voltage is continually reversing, charging, and discharging the capacitor. If the frequency goes to zero, which would be a dc voltage, X_C tends to infinity, and the current is zero once the capacitor is charged. At very high frequencies, the capacitor's reactance tends to zero—it has a negligible reactance and does not impede the current (it acts like a simple wire).

Inductor

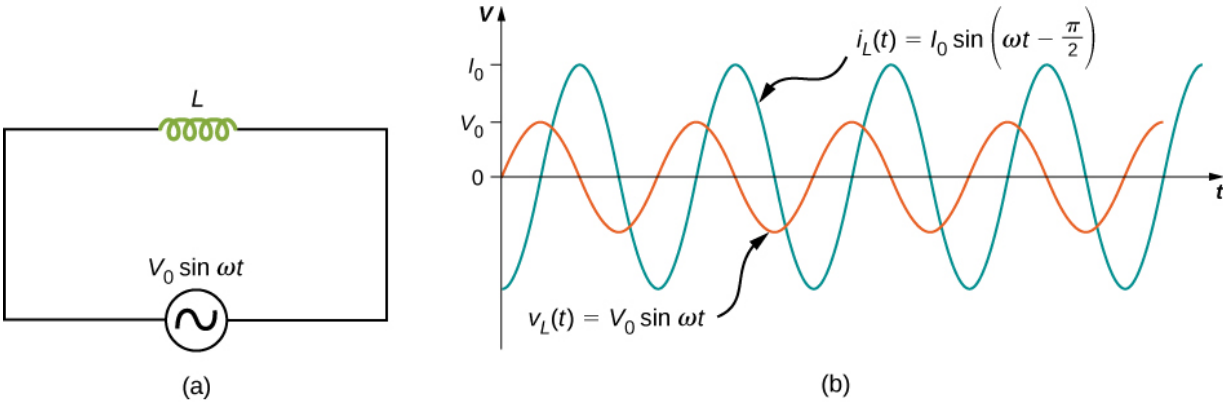
Lastly, let's consider an inductor connected to an ac voltage source. From Kirchhoff's loop rule, the voltage across the inductor L of [link](a) is
Equation:

$$v_L(t) = V_0 \sin \omega t.$$

The emf across an inductor is equal to $\varepsilon = -L\left(di_L/dt\right)$; however, the potential difference across the inductor is $v_L\left(t\right) = Ldi_L\left(t\right)/dt$, because if we consider that the voltage around the loop must equal zero, the voltage gained from the ac source must dissipate through the inductor. Therefore, connecting this with the ac voltage source, we have

Equation:

$$\frac{di_L\left(t\right)}{dt} = \frac{V_0}{L}\sin\omega t.$$


L
$V_0 \sin \omega t$
(a)
v
I_0
V_0
0
t
$i_L(t) = I_0 \sin\left(\omega t - \frac{\pi}{2}\right)$
$v_L(t) = V_0 \sin \omega t$
(b)

(a) An inductor connected across an ac generator. (b) The current $i_L\left(t\right)$ through the inductor and the voltage $v_L\left(t\right)$ across the inductor. Here $i_L\left(t\right)$ lags $v_L\left(t\right)$ by $\pi/2$ rad.

The current $i_L\left(t\right)$ is found by integrating this equation. Since the circuit does not contain a source of constant emf, there is no steady current in the circuit. Hence, we can set the constant of integration, which represents the steady current in the circuit, equal to zero, and we have

Equation:

$$i_L\left(t\right) = -\frac{V_0}{\omega L}\cos\omega t = \frac{V_0}{\omega L}\sin\left(\omega t - \frac{\pi}{2}\right) = I_0 \sin\left(\omega t - \frac{\pi}{2}\right),$$

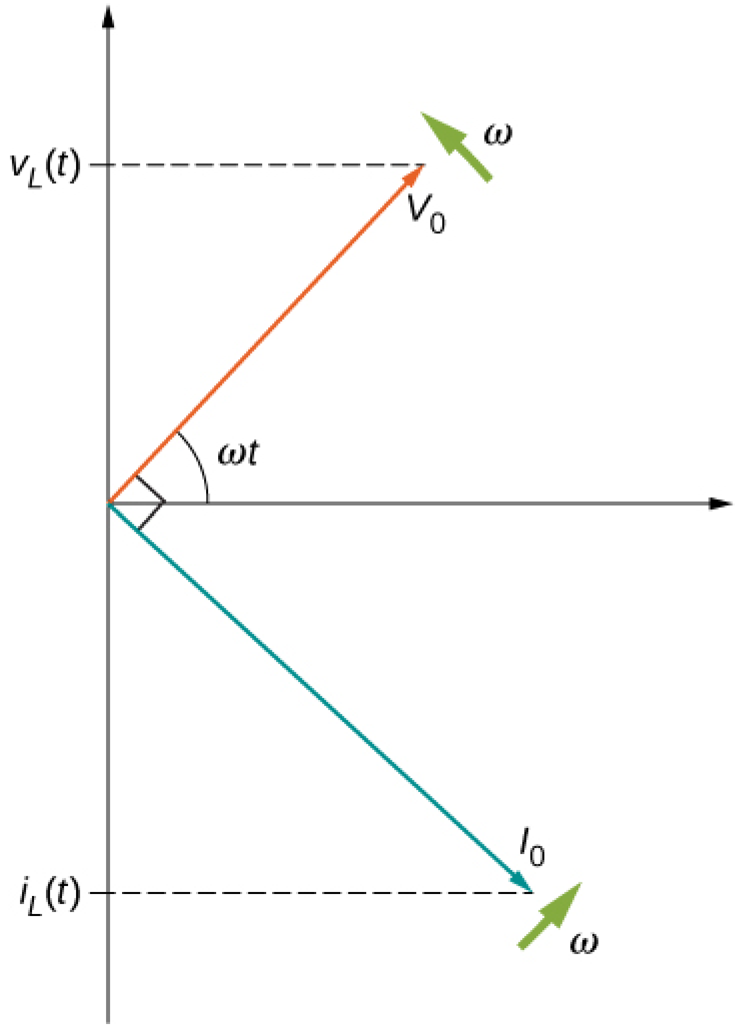
where $I_0 = V_0/\omega L$. The relationship between V_0 and I_0 may also be written in a form analogous to Ohm’s law:

Note:
Equation:

$$\frac{V_0}{I_0} = \omega L = X_L.$$

The quantity X_L is known as the **inductive reactance** of the inductor, or the opposition of an inductor to a change in current; its unit is also the ohm. Note that X_L varies directly as the frequency of the ac source—high frequency causes high inductive reactance.

A phase difference of $\pi/2$ rad occurs between the current through and the voltage across the inductor. From [link] and [link], the current through an inductor lags the potential difference across an inductor by $\pi/2\,\text{rad}$, or a quarter of a cycle. The phasor diagram for this case is shown in [link].



The phasor diagram for the inductor of [link]. The current phasor lags the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

Note:
An animation from the University of New South Wales AC Circuits illustrates some of the concepts we discuss in this chapter. They also include wave and phasor diagrams that evolve over time so that you can get a better picture of how each changes over time.

Example:

Simple AC Circuits

An ac generator produces an emf of amplitude 10 V at a frequency $f=60\,\text{Hz}$. Determine the voltages across and the currents through the circuit elements when the generator is connected to (a) a $100-\Omega$ resistor, (b) a $10-\mu\text{F}$ capacitor, and (c) a 15-mH inductor.

Strategy

The entire AC voltage across each device is the same as the source voltage. We can find the currents by finding the reactance *X* of each device and solving for the peak current using $I_0=V_0/X$.

Solution

The voltage across the terminals of the source is

Equation:

$$v(t)=V_0\sin\omega t=(10\,\text{V})\sin 120\pi t,$$

where $\omega=2\pi f=120\pi\,\text{rad/s}$ is the angular frequency. Since $v(t)$ is also the voltage across each of the elements, we have

Equation:

$$v(t)=v_R(t)=v_C(t)=v_L(t)=(10\,\text{V})\sin 120\pi t.$$

a. When $R=100\,\Omega$, the amplitude of the current through the resistor is

Equation:

$$I_0=V_0/R=10\,\text{V}/100\,\Omega=0.10\,\text{A},$$

so

Equation:

$$i_R(t)=(0.10\,\text{A})\sin 120\pi t.$$

b. From [link], the capacitive reactance is

Equation:

$$X_C=\frac{1}{\omega C}=\frac{1}{(120\pi\,\text{rad/s})(10\,\times\,10^{-6}\text{F})}=265\,\Omega,$$

so the maximum value of the current is

Equation:

$$I_0 = \frac{V_0}{X_C} = \frac{10\text{ V}}{265\,\Omega} = 3.8 \times 10^{-2}\text{ A}$$

and the instantaneous current is given by

Equation:

$$i_C(t) = (3.8 \times 10^{-2}\text{ A})\sin\left(120\pi t + \frac{\pi}{2}\right).$$

c. From [link], the inductive reactance is

Equation:

$$X_L = \omega L = (120\pi\text{ rad/s})(15 \times 10^{-3}\text{ H}) = 5.7\,\Omega.$$

The maximum current is therefore

Equation:

$$I_0 = \frac{10\text{ V}}{5.7\,\Omega} = 1.8\text{ A}$$

and the instantaneous current is

Equation:

$$i_L(t) = (1.8\text{ A})\sin\left(120\pi t - \frac{\pi}{2}\right).$$

Significance

Although the voltage across each device is the same, the peak current has different values, depending on the reactance. The reactance for each device depends on the values of resistance, capacitance, or inductance.

Note:

Exercise:

Problem:

Check Your Understanding Repeat [link] for an ac source of amplitude 20 V and frequency 100 Hz.

Solution:

a. $(20\text{ V})\sin 200\pi t,\ (0.20\text{ A})\sin 200\pi t$; b. $(20\text{ V})\sin 200\pi t,\ (0.13\text{ A})\sin\ (200\pi t+\pi/2)$; c. $(20\text{ V})\sin 200\pi t,\ (2.1\text{ A})\sin\ (200\pi t-\pi/2)$

Summary

- For resistors, the current through and the voltage across are in phase.
- For capacitors, we find that when a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle. Since a capacitor can stop current when fully charged, it limits current and offers another form of ac resistance, called capacitive reactance, which has units of ohms.
- For inductors in ac circuits, we find that when a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle.
- The opposition of an inductor to a change in current is expressed as a type of ac reactance. This inductive reactance, which has units of ohms, varies with the frequency of the ac source.

Conceptual Questions

Exercise:

Problem:

Explain why at high frequencies a capacitor acts as an ac short, whereas an inductor acts as an open circuit.

Problems

Exercise:

Problem:

Calculate the reactance of a $5.0\text{-}\mu\text{F}$ capacitor at (a) 60 Hz, (b) 600 Hz, and (c) 6000 Hz.

Solution:

a. $530\ \Omega$; b. $53\ \Omega$; c. $5.3\ \Omega$

Exercise:

Problem:

What is the capacitance of a capacitor whose reactance is $10\ \Omega$ at 60 Hz?

Exercise:

Problem:

Calculate the reactance of a 5.0-mH inductor at (a) 60 Hz, (b) 600 Hz, and (c) 6000 Hz.

Solution:

a. $1.9\ \Omega$; b. $19\ \Omega$; c. $190\ \Omega$

Exercise:

Problem:

What is the self-inductance of a coil whose reactance is $10\ \Omega$ at 60 Hz?

Exercise:

Problem:

At what frequency is the reactance of a $20\text{-}\mu\text{F}$ capacitor equal to that of a 10-mH inductor?

Solution:

360 Hz

Exercise:

Problem:

At 1000 Hz, the reactance of a 5.0-mH inductor is equal to the reactance of a particular capacitor. What is the capacitance of the capacitor?

Exercise:

Problem:

A $50\text{-}\Omega$ resistor is connected across the emf $v\left(t\right)=\left(160\text{ V}\right)\text{sin}\ \left(120\pi t\right)$. Write an expression for the current through the resistor.

Solution:

$i\left(t\right)=\left(3.2\text{ A}\right)\text{sin}\ \left(120\pi t\right)$

Exercise:

Problem:

A $25\text{-}\mu\text{F}$ capacitor is connected to an emf given by $v\left(t\right)=\left(160\text{ V}\right)\text{sin}\ \left(120\pi t\right)$. (a) What is the reactance of the capacitor? (b) Write an expression for the current output of the source.

Exercise:

Problem:

A 100-mH inductor is connected across the emf of the preceding problem. (a) What is the reactance of the inductor? (b) Write an expression for the current through the inductor.

Solution:

a. $38\ \Omega$; b. $i(t) = (4.24\text{A})\sin\ (120\pi t - \pi/2)$

Glossary

capacitive reactance
opposition of a capacitor to a change in current

inductive reactance
opposition of an inductor to a change in current

rms current
root mean square of the current

rms voltage
root mean square of the voltage

RLC Series Circuits with AC

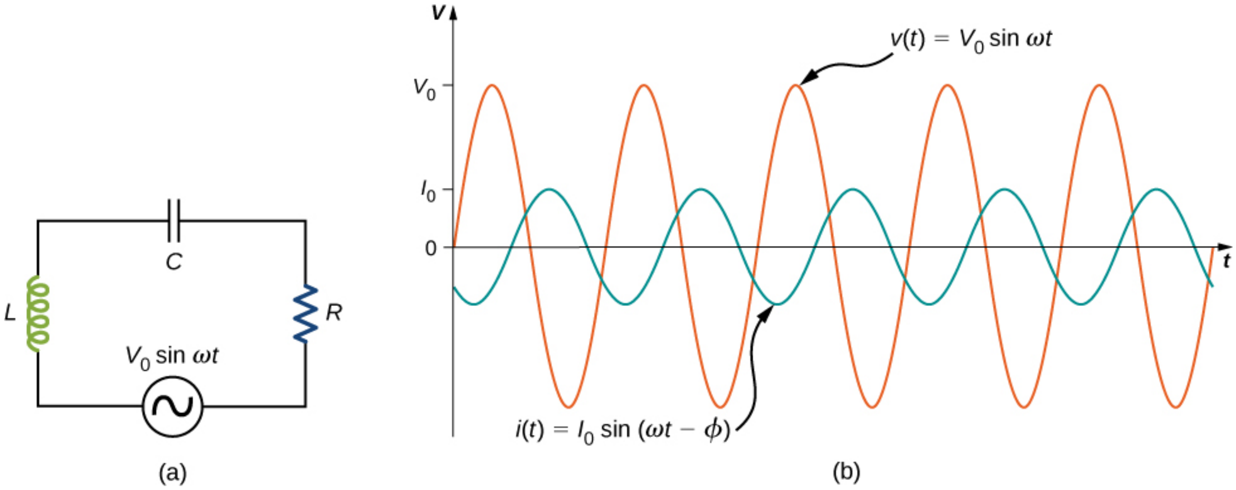
By the end of the section, you will be able to:

- Describe how the current varies in a resistor, a capacitor, and an inductor while in series with an ac power source
- Use phasors to understand the phase angle of a resistor, capacitor, and inductor ac circuit and to understand what that phase angle means
- Calculate the impedance of a circuit

The ac circuit shown in [link], called an *RLC* series circuit, is a series combination of a resistor, capacitor, and inductor connected across an ac source. It produces an emf of

Equation:

$$v(t) = V_0 \sin \omega t.$$



C
L
R
$V_0 \sin \omega t$
(a)
V
$v(t) = V_0 \sin \omega t$
V_0
I_0
0
t
$i(t) = I_0 \sin(\omega t - \phi)$
(b)

(a) An *RLC* series circuit. (b) A comparison of the generator output voltage and the current. The value of the phase difference ϕ depends on the values of *R*, *C*, and *L*.

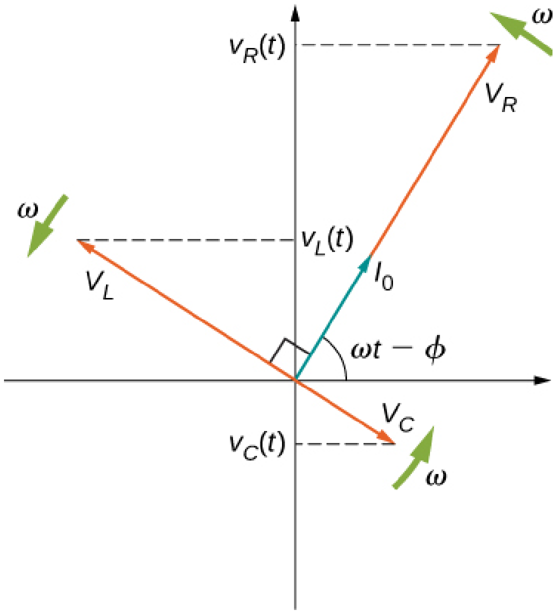
Since the elements are in series, the same current flows through each element at all points in time. The relative phase between the current and the emf is not obvious when all three elements are present. Consequently, we represent the current by the general expression

Equation:

$$i(t) = I_0 \sin(\omega t - \phi),$$

where I_0 is the current amplitude and ϕ is the **phase angle** between the current and the applied voltage. The phase angle is thus the amount by which the voltage and current are out of phase with each other in a circuit. Our task is to find I_0 and ϕ.

A phasor diagram involving $i(t), v_R(t), v_C(t),$ and $v_L(t)$ is helpful for analyzing the circuit. As shown in [link], the phasor representing $v_R(t)$ points in the same direction as the phasor for $i(t)$; its amplitude is $V_R = I_0 R$. The $v_C(t)$ phasor lags the *i*(*t*) phasor by $\pi/2$ rad and has the amplitude $V_C = I_0 X_C$. The phasor for $v_L(t)$ leads the *i*(*t*) phasor by $\pi/2$ rad and has the amplitude $V_L = I_0 X_L$.



The phasor diagram for the *RLC* series circuit of [link].

At any instant, the voltage across the *RLC* combination is $v_R(t) + v_L(t) + v_C(t) = v(t)$, the emf of the source. Since a component of a sum of vectors is the sum of the components of the individual vectors—for example, $(A + B)_y = A_y + B_y$—the projection of the vector sum of phasors onto the vertical axis is the sum of the vertical projections of the individual phasors. Hence, if we add vectorially the phasors representing $v_R(t), v_L(t),$ and $v_C(t)$ and then find the projection of the resultant onto the vertical axis, we obtain

Equation:

$$v_R(t) + v_L(t) + v_C(t) = v(t) = V_0 \sin \omega t.$$

The vector sum of the phasors is shown in [link]. The resultant phasor has an amplitude V_0 and is directed at an angle ϕ with respect to the $v_R(t)$, or $i(t)$, phasor. The projection of this resultant phasor onto the vertical axis is $v(t) = V_0 \sin \omega t$. We can easily determine the unknown quantities I_0 and ϕ from the geometry of the phasor diagram. For the phase angle,

Equation:

$$\phi = \tan^{-1} \frac{V_L - V_C}{V_R} = \tan^{-1} \frac{I_0 X_L - I_0 X_C}{I_0 R},$$

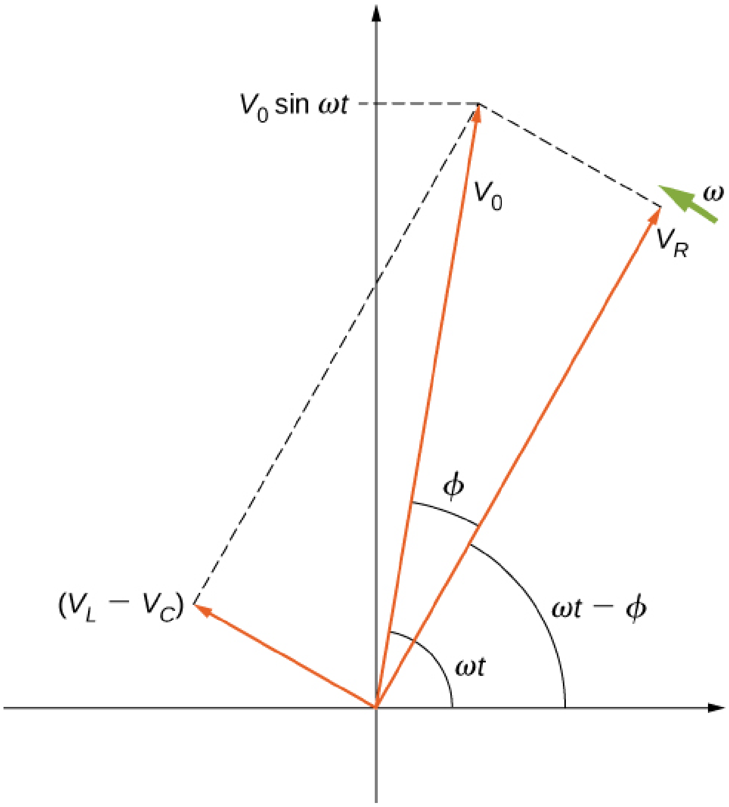
and after cancellation of I_0, this becomes

Note:
Equation:

$$\phi = \tan^{-1}\frac{X_L - X_C}{R}.$$

Furthermore, from the Pythagorean theorem,
Equation:

$$V_0 = \sqrt{V_R^2 + (V_L - V_C)^2} = \sqrt{(I_0R)^2 + (I_0X_L - I_0X_C)^2} = I_0\sqrt{R^2 + (X_L - X_C)^2}.$$


$V_0 \sin \omega t$
V_0
ω
V_R
ϕ
$(V_L - V_C)$
$\omega t - \phi$
ωt

The resultant of the phasors for $v_L(t)$, $v_C(t)$, and $v_R(t)$ is equal to the phasor for $v(t) = V_0 \sin \omega t$. The $i(t)$ phasor (not shown) is aligned with the $v_R(t)$ phasor.

The current amplitude is therefore the ac version of Ohm’s law:

Note:
Equation:

$$I_0 = \frac{V_0}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{V_0}{Z},$$

where

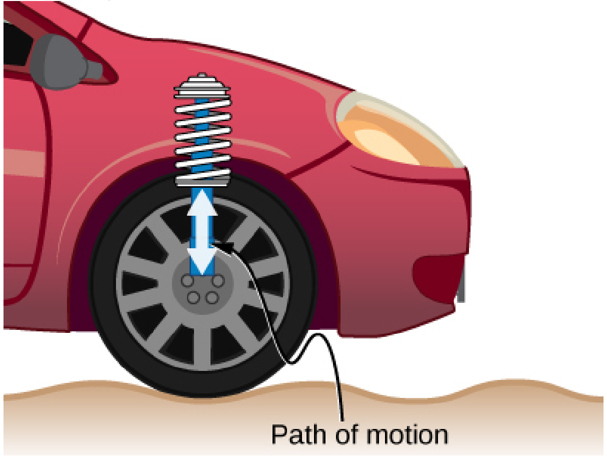
Note:
Equation:

$$Z = \sqrt{R^2 + (X_L - X_C)^2}$$

is known as the **impedance** of the circuit. Its unit is the ohm, and it is the ac analog to resistance in a dc circuit, which measures the combined effect of resistance, capacitive reactance, and inductive reactance ([link]).

Power capacitors are used to balance the impedance of the effective inductance in transmission lines.

The *RLC* circuit is analogous to the wheel of a car driven over a corrugated road ([link]). The regularly spaced bumps in the road drive the wheel up and down; in the same way, a voltage source increases and decreases. The shock absorber acts like the resistance of the *RLC* circuit, damping and limiting the amplitude of the oscillation. Energy within the wheel system goes back and forth between kinetic and potential energy stored in the car spring, analogous to the shift between a maximum current, with energy stored in an inductor, and no current, with energy stored in the electric field of a capacitor. The amplitude of the wheel's motion is at a maximum if the bumps in the road are hit at the resonant frequency, which we describe in more detail in Resonance in an AC Circuit.



On a car, the shock absorber damps motion and dissipates energy. This is much like the resistance in an *RLC* circuit. The mass and spring determine the resonant frequency.

Note:
Problem-Solving Strategy: AC Circuits
To analyze an ac circuit containing resistors, capacitors, and inductors, it is helpful to think of each device's reactance and find the equivalent reactance using the rules we used for equivalent resistance in the past. Phasors are a great method to determine whether the emf of the circuit has positive or negative phase (namely, leads or lags other values). A mnemonic device of "ELI the ICE man" is sometimes used to remember that the emf (E) leads the current (I) in an inductor (L) and the current (I) leads the emf (E) in a capacitor (C).
Use the following steps to determine the emf of the circuit by phasors:

1. Draw the phasors for voltage across each device: resistor, capacitor, and inductor, including the phase angle in the circuit.
2. If there is both a capacitor and an inductor, find the net voltage from these two phasors, since they are antiparallel.

3. Find the equivalent phasor from the phasor in step 2 and the resistor's phasor using trigonometry or components of the phasors. The equivalent phasor found is the emf of the circuit.

Example:

An *RLC* Series Circuit

The output of an ac generator connected to an *RLC* series combination has a frequency of 200 Hz and an amplitude of 0.100 V. If $R = 4.00\,\Omega$, $L = 3.00 \times 10^{-3}\,\text{H}$, and $C = 8.00 \times 10^{-4}\,\text{F}$, what are (a) the capacitive reactance, (b) the inductive reactance, (c) the impedance, (d) the current amplitude, and (e) the phase difference between the current and the emf of the generator?

Strategy

The reactances and impedance in (a)–(c) are found by substitutions into [link], [link], and [link], respectively. The current amplitude is calculated from the peak voltage and the impedance. The phase difference between the current and the emf is calculated by the inverse tangent of the difference between the reactances divided by the resistance.

Solution

a. From [link], the capacitive reactance is
   Equation:

$$X_C = \frac{1}{\omega C} = \frac{1}{2\pi\,(200\,\text{Hz})\left(8.00 \times 10^{-4}\,\text{F}\right)} = 0.995\,\Omega.$$

b. From [link], the inductive reactance is
   Equation:

$$X_L = \omega L = 2\pi\,(200\,\text{Hz})\left(3.00 \times 10^{-3}\,\text{H}\right) = 3.77\,\Omega.$$

c. Substituting the values of R, X_C, and X_L into [link], we obtain for the impedance
   Equation:

$$Z = \sqrt{(4.00\,\Omega)^2 + (3.77\,\Omega - 0.995\,\Omega)^2} = 4.87\,\Omega.$$

d. The current amplitude is
   Equation:

$$I_0 = \frac{V_0}{Z} = \frac{0.100\,\text{V}}{4.87\,\Omega} = 2.05 \times 10^{-2}\text{A}.$$

e. From [link], the phase difference between the current and the emf is
   Equation:

$$\phi = \tan^{-1}\frac{X_L - X_C}{R} = \tan^{-1}\frac{2.77\,\Omega}{4.00\,\Omega} = 0.607\,\text{rad}.$$

Significance

The phase angle is positive because the reactance of the inductor is larger than the reactance of the capacitor.

Note:
Exercise:

Problem:

Check Your Understanding Find the voltages across the resistor, the capacitor, and the inductor in the circuit of [link] using $v(t) = V_0 \sin \omega t$ as the output of the ac generator.

Solution:

$v_R = (V_0 R/Z) \sin\ (\omega t - \phi); v_C = (V_0 X_C/Z) \sin\ (\omega t - \phi + \pi/2) = -(V_0 X_C/Z) \cos(\omega t - \phi);$
$v_L = (V_0 X_L/Z) \sin\ (\omega t - \phi - \pi/2) = (V_0 X_L/Z) \cos(\omega t - \phi)$

Summary

- An *RLC* series circuit is a resistor, capacitor, and inductor series combination across an ac source.
- The same current flows through each element of an *RLC* series circuit at all points in time.
- The counterpart of resistance in a dc circuit is impedance, which measures the combined effect of resistors, capacitors, and inductors. The maximum current is defined by the ac version of Ohm's law.
- Impedance has units of ohms and is found using the resistance, the capacitive reactance, and the inductive reactance.

Conceptual Questions

Exercise:

Problem:

In an *RLC* series circuit, can the voltage measured across the capacitor be greater than the voltage of the source? Answer the same question for the voltage across the inductor.

Solution:

yes for both

Problems

Exercise:

Problem:

What is the impedance of a series combination of a $50\text{-}\Omega$ resistor, a $5.0\text{-}\mu\text{F}$ capacitor, and a $10\text{-}\mu\text{F}$ capacitor at a frequency of 2.0 kHz?

Exercise:

Problem:

A resistor and capacitor are connected in series across an ac generator. The emf of the generator is given by $v(t) = V_0 \cos \omega t$, where $V_0 = 120 \text{ V}$, $\omega = 120\pi \text{ rad/s}$, $R = 400\,\Omega$, and $C = 4.0\mu\text{F}$. (a) What is the impedance of the circuit? (b) What is the amplitude of the current through the resistor? (c) Write an expression for the current through the resistor. (d) Write expressions representing the voltages across the resistor and across the capacitor.

Solution:

a. $770\,\Omega$; b. 0.16 A; c. $I = (0.16 \text{ A})\cos(120\pi t)$; d. $v_R = 62\cos(120\pi t)$; $v_C = 103\cos(120\pi t - \pi/2)$

Exercise:

Problem:

A resistor and inductor are connected in series across an ac generator. The emf of the generator is given by $v(t) = V_0 \cos \omega t$, where $V_0 = 120 \text{ V}$ and $\omega = 120\pi \text{ rad/s}$; also, $R = 400\,\Omega$ and $L = 1.5 \text{ H}$. (a) What is the impedance of the circuit? (b) What is the amplitude of the current through the resistor? (c) Write an expression for the current through the resistor. (d) Write expressions representing the voltages across the resistor and across the inductor.

Exercise:

Problem:

In an *RLC* series circuit, the voltage amplitude and frequency of the source are 100 V and 500 Hz, respectively, an $R = 500\,\Omega$, $L = 0.20 \text{ H}$, and $C = 2.0\mu\text{F}$. (a) What is the impedance of the circuit? (b) What is the amplitude of the current from the source? (c) If the emf of the source is given by $v(t) = (100 \text{ V}) \sin 1000\pi t$, how does the current vary with time? (d) Repeat the calculations with *C* changed to $0.20\mu\text{F}$.

Solution:

a. $690\,\Omega$; b. 0.15 A; c. $I = (0.15\text{A}) \sin\,(1000\pi t - 0.753)$; d. $1100\,\Omega$, 0.092 A, $I = (0.092\text{A}) \sin\,(1000\pi t + 1.09)$

Exercise:

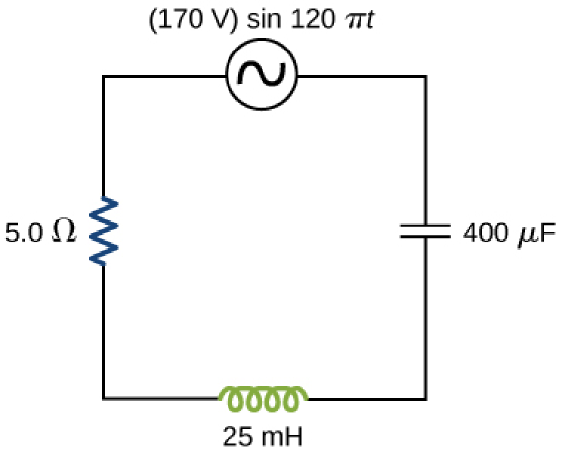
Problem:

An *RLC* series circuit with $R = 600\,\Omega$, $L = 30 \text{ mH}$, and $C = 0.050\mu\text{F}$ is driven by an ac source whose frequency and voltage amplitude are 500 Hz and 50 V, respectively. (a) What is the impedance of the circuit? (b) What is the amplitude of the current in the circuit? (c) What is the phase angle between the emf of the source and the current?

Exercise:

Problem:

For the circuit shown below, what are (a) the total impedance and (b) the phase angle between the current and the emf? (c) Write an expression for $i(t)$.


(170 V) sin 120 πt
5.0 Ω
400 μF
25 mH

Solution:

a. $5.7\ \Omega$; b. 29°; c. $I = (30.\ \mathrm{A})\cos(120\pi t + 0.51)$

Glossary

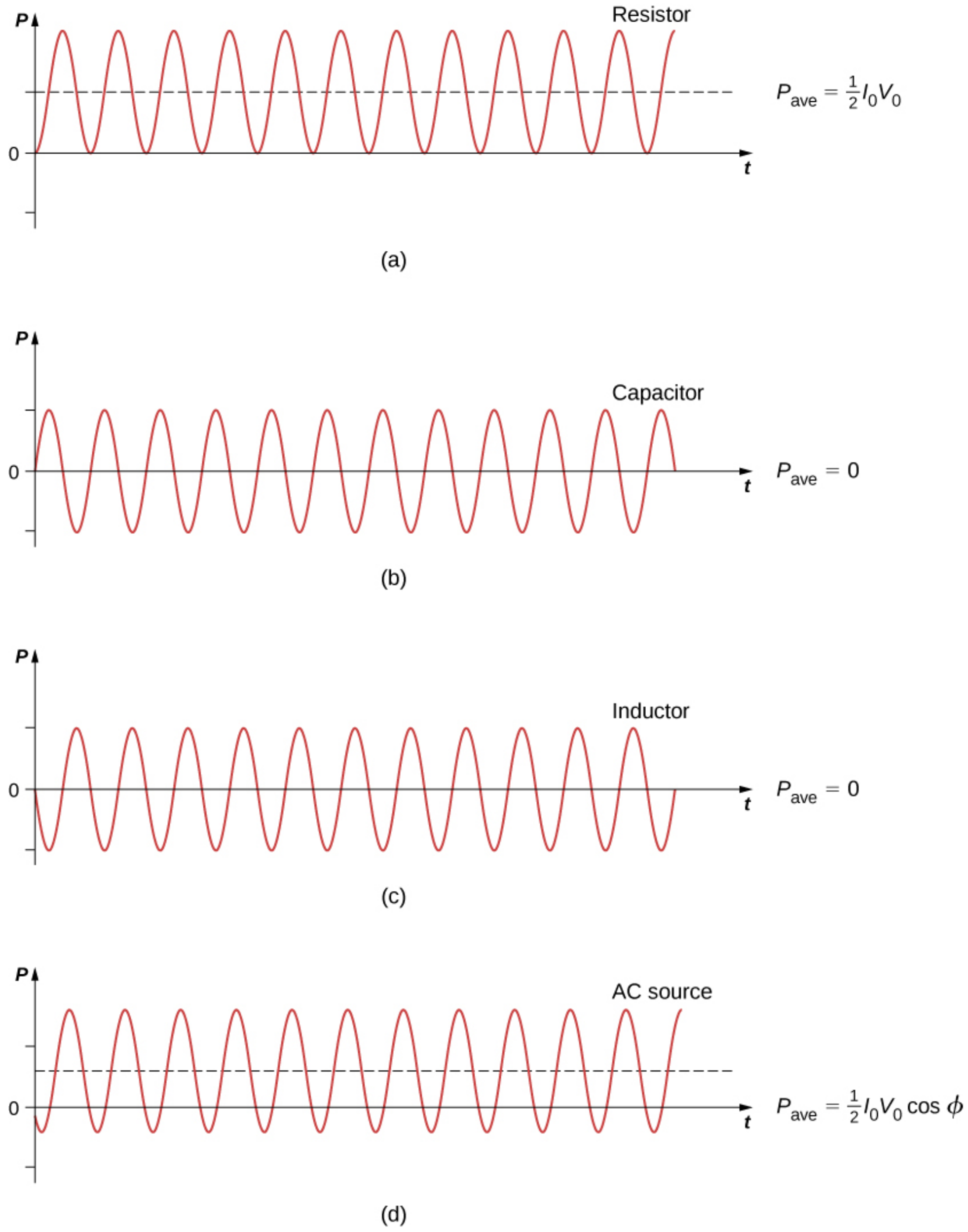
impedance	ac analog to resistance in a dc circuit, which measures the combined effect of resistance, capacitive reactance, and inductive reactance
phase angle	amount by which the voltage and current are out of phase with each other in a circuit

Power in an AC Circuit

By the end of the section, you will be able to:

- Describe how average power from an ac circuit can be written in terms of peak current and voltage and of rms current and voltage
- Determine the relationship between the phase angle of the current and voltage and the average power, known as the power factor

A circuit element dissipates or produces power according to $P=IV$, where I is the current through the element and V is the voltage across it. Since the current and the voltage both depend on time in an ac circuit, the instantaneous power $p(t)=i(t)v(t)$ is also time dependent. A plot of $p(t)$ for various circuit elements is shown in [link]. For a resistor, $i(t)$ and $v(t)$ are in phase and therefore always have the same sign (see [link]). For a capacitor or inductor, the relative signs of $i(t)$ and $v(t)$ vary over a cycle due to their phase differences (see [link] and [link]). Consequently, $p(t)$ is positive at some times and negative at others, indicating that capacitive and inductive elements produce power at some instants and absorb it at others.



Graph of instantaneous power for various circuit elements. (a) For the resistor, $P_{\text{ave}} = I_0 V_0 / 2$, whereas for (b) the capacitor and (c) the inductor, $P_{\text{ave}} = 0$. (d) For the source, $P_{\text{ave}} = I_0 V_0 (\cos \phi)/2$, which may be positive, negative, or zero, depending on ϕ.

Because instantaneous power varies in both magnitude and sign over a cycle, it seldom has any practical importance. What we’re almost always concerned with is the power averaged over time, which we refer to as the **average power**. It is defined by the time average of the instantaneous power over one cycle:

Equation:

$$P_{\text{ave}} = \frac{1}{T}\int_0^T p(t)dt,$$

where $T = 2\pi/\omega$ is the period of the oscillations. With the substitutions $v(t) = V_0 \sin \omega t$ and $i(t) = I_0 \sin(\omega t - \phi)$, this integral becomes

Equation:

$$P_{\text{ave}} = \frac{I_0 V_0}{T}\int_0^T \sin(\omega t - \phi)\sin \omega t \, dt.$$

Using the trigonometric relation $\sin(A - B) = \sin A \cos B - \sin B \cos A$, we obtain

Equation:

$$P_{\text{ave}} = \frac{I_0 V_0 \cos \phi}{T}\int_0^T \sin^2 \omega t dt - \frac{I_0 V_0 \sin \phi}{T}\int_0^T \sin \omega t \cos \omega t dt.$$

Evaluation of these two integrals yields

Equation:

$$\frac{1}{T}\int_0^T \sin^2 \omega t dt = \frac{1}{2}$$

and

Equation:

$$\frac{1}{T}\int_0^T \sin \omega t \cos \omega t dt = 0.$$

Hence, the average power associated with a circuit element is given by

Note:

Equation:

$$P_{\text{ave}} = \frac{1}{2} I_0 V_0 \cos \phi.$$

In engineering applications, $\cos \phi$ is known as the **power factor**, which is the amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase. For a resistor, $\phi = 0$, so the average power dissipated is

Equation:

$$P_{\text{ave}} = \frac{1}{2} I_0 V_0.$$

A comparison of *p*(*t*) and P_{ave} is shown in [link](d). To make $P_{\text{ave}} = (1/2)I_0V_0$ look like its dc counterpart, we use the rms values I_{rms} and V_{rms} of the current and the voltage. By definition, these are

Equation:

$$I_{\text{rms}} = \sqrt{i^2_{\text{ave}}} \text{ and } V_{\text{rms}} = \sqrt{v^2_{\text{ave}}},$$

where

Equation:

$$i^2_{\text{ave}} = \frac{1}{T}\int_0^T i^2(t)dt \text{ and } v^2_{\text{ave}} = \frac{1}{T}\int_0^T v^2(t)dt.$$

With $i(t) = I_0 \sin(\omega t - \phi)$ and $v(t) = V_0 \sin \omega t$, we obtain

Equation:

$$I_{\text{rms}} = \frac{1}{\sqrt{2}}I_0 \text{ and } V_{\text{rms}} = \frac{1}{\sqrt{2}}V_0.$$

We may then write for the average power dissipated by a resistor,

Note:

Equation:

$$P_{\text{ave}} = \frac{1}{2}I_0V_0 = I_{\text{rms}}V_{\text{rms}} = I^2_{\text{rms}}R.$$

This equation further emphasizes why the rms value is chosen in discussion rather than peak values. Both equations for average power are correct for [link], but the rms values in the formula give a cleaner representation, so the extra factor of 1/2 is not necessary.

Alternating voltages and currents are usually described in terms of their rms values. For example, the 110 V from a household outlet is an rms value. The amplitude of this source is $110\sqrt{2}\,\text{V} = 156\text{ V}$. Because most ac meters are calibrated in terms of rms values, a typical ac voltmeter placed across a household outlet will read 110 V.

For a capacitor and an inductor, $\phi = \pi/2$ and $-\pi/2$ rad, respectively. Since $\cos\pi/2 = \cos(-\pi/2) = 0$, we find from [link] that the average power dissipated by either of these elements is $P_{\text{ave}} = 0$. Capacitors and inductors absorb energy from the circuit during one half-cycle and then discharge it back to the circuit during the other half-cycle. This behavior is illustrated in the plots of [link], (b) and (c), which show *p*(t) oscillating sinusoidally about zero.

The phase angle for an ac generator may have any value. If $\cos\phi > 0$, the generator produces power; if $\cos\phi < 0$, it absorbs power. In terms of rms values, the average power of an ac generator is written as

Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos \phi .$$

For the generator in an *RLC* circuit,
Equation:

$$\tan \phi = \frac{X_L - X_C}{R}$$

and
Equation:

$$\cos \phi = \frac{R}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{R}{Z} .$$

Hence the average power of the generator is
Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos \phi = \frac{V_{\text{rms}}}{Z} V_{\text{rms}} \frac{R}{Z} = \frac{V_{\text{rms}}^2 R}{Z^2} .$$

This can also be written as
Equation:

$$P_{\text{ave}} = I_{\text{rms}}^2 R,$$

which designates that the power produced by the generator is dissipated in the resistor. As we can see, Ohm’s law for the rms ac is found by dividing the rms voltage by the impedance.

Example:
Power Output of a Generator
An ac generator whose emf is given by
Equation:

$$v(t) = (4.00 \text{ V}) \sin \left[\left(1.00 \times 10^4 \text{rad/s} \right) t \right]$$

is connected to an *RLC* circuit for which $L = 2.00 \times 10^{-3}$ H, $C = 4.00 \times 10^{-6}$ F, and $R = 5.00 \, \Omega$. (a) What is the rms voltage across the generator? (b) What is the impedance of the circuit? (c) What is the average power output of the generator?
Strategy
The rms voltage is the amplitude of the voltage times $1/\sqrt{2}$. The impedance of the circuit involves the resistance and the reactances of the capacitor and the inductor. The average power is calculated by [link], or more specifically, the last part of the equation, because we have the impedance of the circuit *Z*, the rms voltage V_{rms}, and the resistance *R*.
Solution

a. Since $V_0 = 4.00$ V, the rms voltage across the generator is

Equation:

$$V_{\text{rms}} = \frac{1}{\sqrt{2}}(4.00\ \text{V}) = 2.83\ \text{V}.$$

b. The impedance of the circuit is
Equation:

$$\begin{aligned} Z &= \sqrt{R^2 + (X_L - X_C)^2} \\ &= \left\{(5.00\ \Omega)^2 + \left[\left(1.00\ \times\ 10^4\ \text{rad/s}\right)\left(2.00\ \times\ 10^{-3}\ \text{H}\right) - \frac{1}{\left(1.00 \times 10^4\ \text{rad/s}\right)\left(4.00 \times 10^{-6}\ \text{F}\right)}\right]^2\right\}^{1/2} \\ &= 7.07\ \Omega. \end{aligned}$$

c. From [link], the average power transferred to the circuit is
Equation:

$$P_{\text{ave}} = \frac{V_{\text{rms}}^2 R}{Z^2} = \frac{(2.83\ \text{V})^2\,(5.00\ \Omega)}{(7.07\ \Omega)^2} = 0.801\ \text{W}.$$

Significance
If the resistance is much larger than the reactance of the capacitor or inductor, the average power is a dc circuit equation of $P = V^2/R$, where V replaces the rms voltage.

Note:
Exercise:

Problem:

Check Your Understanding An ac voltmeter attached across the terminals of a 45-Hz ac generator reads 7.07 V. Write an expression for the emf of the generator.

Solution:

$v(t) = (10.0\ \text{V})\ \sin 90\pi t$

Note:
Exercise:

Problem:

Check Your Understanding Show that the rms voltages across a resistor, a capacitor, and an inductor in an ac circuit where the rms current is I_{rms} are given by $I_{\text{rms}}R$, $I_{\text{rms}}X_C$, and $I_{\text{rms}}X_L$, respectively. Determine these values for the components of the *RLC* circuit of [link].

Solution:

2.00 V; 10.01 V; 8.01 V

Summary

- The average ac power is found by multiplying the rms values of current and voltage.
- Ohm’s law for the rms ac is found by dividing the rms voltage by the impedance.
- In an ac circuit, there is a phase angle between the source voltage and the current, which can be found by dividing the resistance by the impedance.
- The average power delivered to an *RLC* circuit is affected by the phase angle.
- The power factor ranges from –1 to 1.

Conceptual Questions

Exercise:

Problem:

For what value of the phase angle ϕ between the voltage output of an ac source and the current is the average power output of the source a maximum?

Exercise:

Problem: Discuss the differences between average power and instantaneous power.

Solution:

The instantaneous power is the power at a given instant. The average power is the power averaged over a cycle or number of cycles.

Exercise:

Problem:

The average ac current delivered to a circuit is zero. Despite this, power is dissipated in the circuit. Explain.

Exercise:

Problem:

Can the instantaneous power output of an ac source ever be negative? Can the average power output be negative?

Solution:

The instantaneous power can be negative, but the power output can’t be negative.

Exercise:

Problem:

The power rating of a resistor used in ac circuits refers to the maximum average power dissipated in the resistor. How does this compare with the maximum instantaneous power dissipated in the resistor?

Problems

Exercise:

Problem:

The emf of an ac source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100$ V and $\omega = 200\pi \text{ rad/s}$. Calculate the average power output of the source if it is connected across (a) a $20\text{-}\mu\text{F}$ capacitor, (b) a 20-mH inductor, and (c) a $50\text{-}\Omega$ resistor.

Exercise:

Problem:

Calculate the rms currents for an ac source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100$ V and $\omega = 200\pi \text{ rad/s}$ when connected across (a) a $20\text{-}\mu\text{F}$ capacitor, (b) a 20-mH inductor, and (c) a $50\text{-}\Omega$ resistor.

Solution:

a. 0.89 A; b. 5.6A; c. 1.4 A

Exercise:

Problem:

A 40-mH inductor is connected to a 60-Hz AC source whose voltage amplitude is 50 V. If an AC voltmeter is placed across the inductor, what does it read?

Exercise:

Problem:

For an *RLC* series circuit, the voltage amplitude and frequency of the source are 100 V and 500 Hz, respectively; $R = 500\ \Omega$; and $L = 0.20\ \text{H}$. Find the average power dissipated in the resistor for the following values for the capacitance: (a) $C = 2.0 \mu\text{F}$ and (b) $C = 0.20\ \mu\text{F}$.

Solution:

a. 5.3 W; b. 2.1 W

Exercise:

Problem:

An ac source of voltage amplitude 10 V delivers electric energy at a rate of 0.80 W when its current output is 2.5 A. What is the phase angle ϕ between the emf and the current?

Exercise:

Problem:

An *RLC* series circuit has an impedance of 60 Ω and a power factor of 0.50, with the voltage lagging the current. (a) Should a capacitor or an inductor be placed in series with the elements to raise the power factor of the circuit? (b) What is the value of the reactance across the inductor that will raise the power factor to unity?

Solution:

a. inductor; b. $X_L = 52\,\Omega$

Glossary

average power
time average of the instantaneous power over one cycle

power factor
amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase

Resonance in an AC Circuit

By the end of the section, you will be able to:

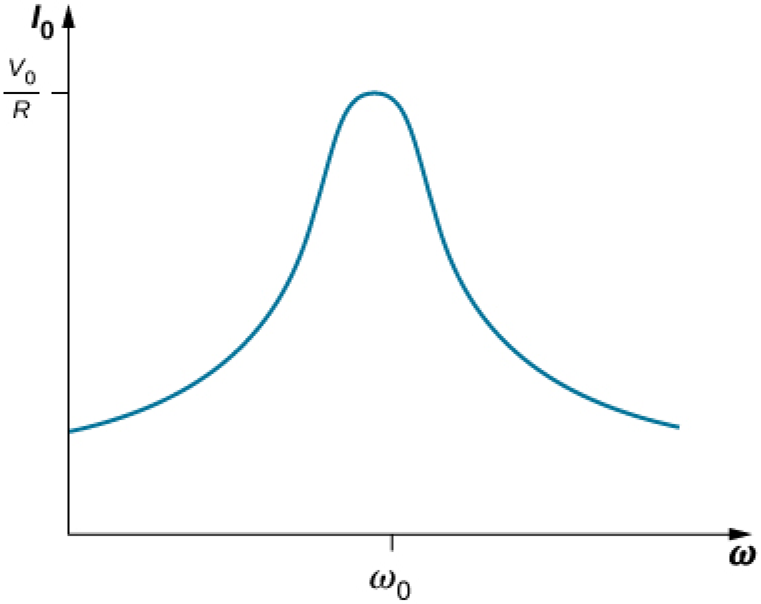
- Determine the peak ac resonant angular frequency for a RLC circuit
- Explain the width of the average power versus angular frequency curve and its significance using terms like bandwidth and quality factor

In the *RLC* series circuit of [link], the current amplitude is, from [link],

Equation:

$$I_0 = \frac{V_0}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}}.$$

If we can vary the frequency of the ac generator while keeping the amplitude of its output voltage constant, then the current changes accordingly. A plot of I_0 versus ω is shown in [link].



At an *RLC* circuit's resonant frequency, $\omega_0 = \sqrt{1/LC}$, the current amplitude is at its maximum value.

In Oscillations, we encountered a similar graph where the amplitude of a damped harmonic oscillator was plotted against the angular frequency of a sinusoidal driving force (see Forced Oscillations). This similarity is more than just a coincidence, as shown earlier by the application of Kirchhoff’s loop rule to the circuit of [link]. This yields

Equation:

$$L\frac{di}{dt}+iR+\frac{q}{C}=V_0\sin\omega t,$$

or

Equation:

$$L\frac{d^2q}{dt^2}+R\frac{dq}{dt}+\frac{1}{C}q=V_0\sin\omega t,$$

where we substituted $dq(\mathrm{t})/dt$ for $i(\mathrm{t})$. A comparison of [link] and, from Oscillations, Damped Oscillations for damped harmonic motion clearly demonstrates that the driven *RLC* series circuit is the electrical analog of the driven damped harmonic oscillator.

The **resonant frequency** f_0 of the *RLC* circuit is the frequency at which the amplitude of the current is a maximum and the circuit would oscillate if not driven by a voltage source. By inspection, this corresponds to the angular frequency $\omega_0=2\pi f_0$ at which the impedance *Z* in [link] is a minimum, or when

Equation:

$$\omega_0 L=\frac{1}{\omega_0 C}$$

and

Note:
Equation:

$$\omega_0 = \sqrt{\frac{1}{LC}}.$$

This is the resonant angular frequency of the circuit. Substituting ω_0 into [link], [link], and [link], we find that at resonance,

Equation:

$$\phi = \tan^{-1}(0) = 0, \quad I_0 = V_0/R, \quad \text{and} \quad Z = R.$$

Therefore, at resonance, an *RLC* circuit is purely resistive, with the applied emf and current in phase.

What happens to the power at resonance? [link] tells us how the average power transferred from an ac generator to the *RLC* combination varies with frequency. In addition, P_{ave} reaches a maximum when Z, which depends on the frequency, is a minimum, that is, when $X_L = X_C$ and $Z = R$. Thus, at resonance, the average power output of the source in an *RLC* series circuit is a maximum. From [link], this maximum is V_{rms}^2/R.

[link] is a typical plot of P_{ave} versus ω in the region of maximum power output. The **bandwidth** Δω of the resonance peak is defined as the range of angular frequencies ω over which the average power P_{ave} is greater than one-half the maximum value of P_{ave}. The sharpness of the peak is described by a dimensionless quantity known as the **quality factor** Q of the circuit. By definition,

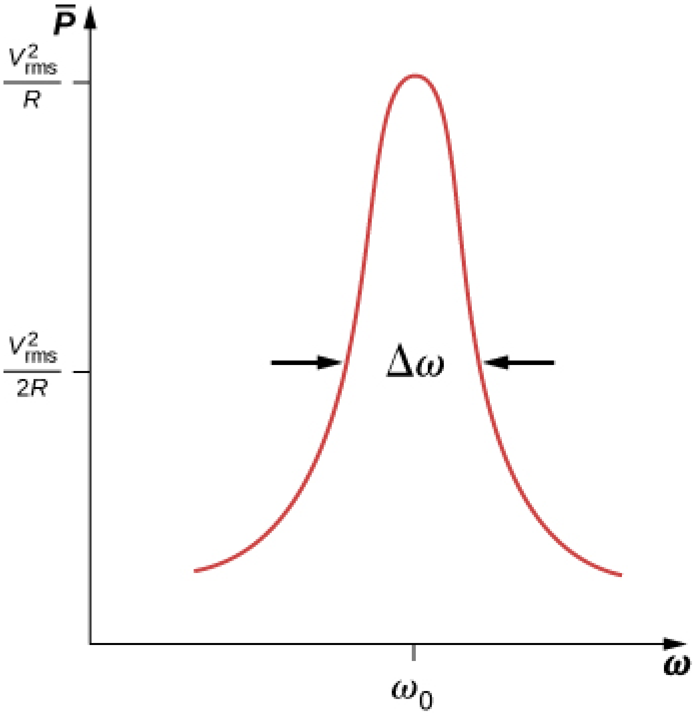
Note:
Equation:

$$Q = \frac{\omega_0}{\Delta\omega},$$

where ω_0 is the resonant angular frequency. A high Q indicates a sharp resonance peak. We can give Q in terms of the circuit parameters as

Note:
Equation:

$$Q = \frac{\omega_0 L}{R}.$$



Like the current, the average power

transferred from an ac generator to an *RLC* circuit peaks at the resonant frequency.

Resonant circuits are commonly used to pass or reject selected frequency ranges. This is done by adjusting the value of one of the elements and hence “tuning” the circuit to a particular resonant frequency. For example, in radios, the receiver is tuned to the desired station by adjusting the resonant frequency of its circuitry to match the frequency of the station. If the tuning circuit has a high Q, it will have a small bandwidth, so signals from other stations at frequencies even slightly different from the resonant frequency encounter a high impedance and are not passed by the circuit. Cell phones work in a similar fashion, communicating with signals of around 1 GHz that are tuned by an inductor-capacitor circuit. One of the most common applications of capacitors is their use in ac-timing circuits, based on attaining a resonant frequency. A metal detector also uses a shift in resonance frequency in detecting metals ([link]).

When a metal detector comes near a piece of metal, the self-inductance of one of its coils changes. This causes a shift in the resonant frequency of a circuit containing the coil. That shift is detected by the circuitry and transmitted to the diver by means of the headphones. (credit: modification of work by Eric Lippmann, U.S. Navy)

Example:

Resonance in an *RLC* Series Circuit

(a) What is the resonant frequency of a circuit using the voltage and LRC values all wired in series from [link]? (b) If the ac generator is set to this frequency without changing the amplitude of the output voltage, what is the amplitude of the current?

Strategy

The resonant frequency for a *RLC* circuit is calculated from [link], which comes from a balance between the reactances of the capacitor and the inductor. Since the circuit is at resonance, the impedance is equal to the resistor. Then, the peak current is calculated by the voltage divided by the resistance.

Solution

a. The resonant frequency is found from [link]:
   Equation:

$$\begin{aligned} f_0 &= \frac{1}{2\pi}\sqrt{\frac{1}{LC}} = \frac{1}{2\pi}\sqrt{\frac{1}{\left(3.00\times10^{-3}\,\text{H}\right)\left(8.00\times10^{-4}\,\text{F}\right)}} \\ &= 1.03 \times 10^{2}\,\text{Hz}. \end{aligned}$$

b. At resonance, the impedance of the circuit is purely resistive, and the current amplitude is
   Equation:

$$I_0 = \frac{0.100\,\text{V}}{4.00\,\Omega} = 2.50 \times 10^{-2}\text{A}.$$

Significance

If the circuit were not set to the resonant frequency, we would need the impedance of the entire circuit to calculate the current.

Example:

Power Transfer in an *RLC* Series Circuit at Resonance

(a) What is the resonant angular frequency of an *RLC* circuit with $R = 0.200\,\Omega$, $L = 4.00 \times 10^{-3}\,\text{H}$, and $C = 2.00 \times 10^{-6}\,\text{F}$? (b) If an ac source of constant amplitude 4.00 V is set to this frequency, what is the average power transferred to the circuit? (c) Determine Q and the bandwidth of this circuit.

Strategy

The resonant angular frequency is calculated from [link]. The average power is calculated from the rms voltage and the resistance in the circuit. The quality factor is calculated from [link] and by knowing the resonant frequency. The bandwidth is calculated from [link] and by knowing the quality factor.

Solution

a. The resonant angular frequency is
   Equation:

$$\begin{aligned}\omega_0 &= \sqrt{\frac{1}{LC}} = \sqrt{\frac{1}{\left(4.00 \times 10^{-3}\,\text{H}\right)\left(2.00 \times 10^{-6}\,\text{F}\right)}} \\ &= 1.12 \times 10^4\,\text{rad/s}.\end{aligned}$$

b. At this frequency, the average power transferred to the circuit is a maximum. It is
   Equation:

$$P_{\text{ave}} = \frac{V_{\text{rms}}^2}{R} = \frac{\left[\left(1/\sqrt{2}\right)(4.00\,\text{V})\right]^2}{0.200\,\Omega} = 40.0\,\text{W}.$$

c. The quality factor of the circuit is
   Equation:

$$Q = \frac{\omega_0 L}{R} = \frac{\left(1.12 \times 10^4\,\text{rad/s}\right)\left(4.00 \times 10^{-3}\,\text{H}\right)}{0.200\,\Omega} = 224.$$

We then find for the bandwidth

Equation:

$$\Delta\omega = \frac{\omega_0}{Q} = \frac{1.12 \times 10^4\,\text{rad/s}}{224} = 50.0\,\text{rad/s}.$$

Significance
If a narrower bandwidth is desired, a lower resistance or higher inductance would help. However, a lower resistance increases the power transferred to the circuit, which may not be desirable, depending on the maximum power that could possibly be transferred.

Note:
Exercise:

Problem:

Check Your Understanding In the circuit of [link], $L = 2.0 \times 10^{-3}\,\text{H}$, $C = 5.0 \times 10^{-4}\,\text{F}$, and $R = 40\,\Omega$. (a) What is the resonant frequency? (b) What is the impedance of the circuit at resonance? (c) If the voltage amplitude is 10 V, what is $i(t)$ at resonance? (d) The frequency of the AC generator is now changed to 200 Hz. Calculate the phase difference between the current and the emf of the generator.

Solution:

a. 160 Hz; b. 40 Ω; c. $(0.25\,\text{A})\sin 10^3 t$; d. 0.023 rad

Note:
Exercise:

Problem:

Check Your Understanding What happens to the resonant frequency of an *RLC* series circuit when the following quantities are increased by a factor of 4: (a) the capacitance, (b) the self-inductance, and (c) the resistance?

Solution:

a. halved; b. halved; c. same

Note:
Exercise:

Problem:

Check Your Understanding The resonant angular frequency of an *RLC* series circuit is $4.0 \times 10^2\,\text{rad/s}$. An ac source operating at this frequency transfers an average power of $2.0 \times 10^{-2}\,\text{W}$ to the circuit. The resistance of the circuit is $0.50\,\Omega$. Write an expression for the emf of the source.

Solution:

$v(t) = (0.14\,\text{V})\,\sin\,(4.0 \times 10^2 t)$

Summary

- At the resonant frequency, inductive reactance equals capacitive reactance.
- The average power versus angular frequency plot for a *RLC* circuit has a peak located at the resonant frequency; the sharpness or width of the peak is known as the bandwidth.

- The bandwidth is related to a dimensionless quantity called the quality factor. A high quality factor value is a sharp or narrow peak.

Problems

Exercise:

Problem:

(a) Calculate the resonant angular frequency of an *RLC* series circuit for which $R = 20\,\Omega$, $L = 75\text{ mH}$, and $C = 4.0\mu\text{F}.$ (b) If *R* is changed to $300\,\Omega$, what happens to the resonant angular frequency?

Exercise:

Problem:

The resonant frequency of an *RLC* series circuit is $2.0\ \times\ 10^3\text{Hz}.$ If the self-inductance in the circuit is 5.0 mH, what is the capacitance in the circuit?

Solution:

$1.3\ \times\ 10^{-6}\text{ F}$

Exercise:

Problem:

(a) What is the resonant frequency of an *RLC* series circuit with $R = 20\,\Omega$, $L = 2.0\,\text{mH}$, and $C = 4.0\mu\text{F}$? (b) What is the impedance of the circuit at resonance?

Exercise:

Problem:

For an *RLC* series circuit, $R = 100\,\Omega$, $L = 150\,\text{mH}$, and $C = 0.25\mu\text{F}.$ (a) If an ac source of variable frequency is connected to the circuit, at what frequency is maximum power dissipated in the resistor? (b) What is the quality factor of the circuit?

Solution:

a. 820 Hz; b. 7.8

Exercise:

Problem:

An ac source of voltage amplitude 100 V and variable frequency *f* drives an *RLC* series circuit with $R = 10\,\Omega$, $L = 2.0\,\text{mH}$, and $C = 25\mu\text{F}$. (a) Plot the current through the resistor as a function of the frequency *f*. (b) Use the plot to determine the resonant frequency of the circuit.

Exercise:

Problem:

(a) What is the resonant frequency of a resistor, capacitor, and inductor connected in series if $R = 100\,\Omega$, $L = 2.0\,\text{H}$, and $C = 5.0\mu\text{F}$? (b) If this combination is connected to a 100-V source operating at the resonant frequency, what is the power output of the source? (c) What is the *Q* of the circuit? (d) What is the bandwidth of the circuit?

Solution:

a. 50 Hz; b. 50 W; c. 6.32; d. 50 rad/s

Exercise:

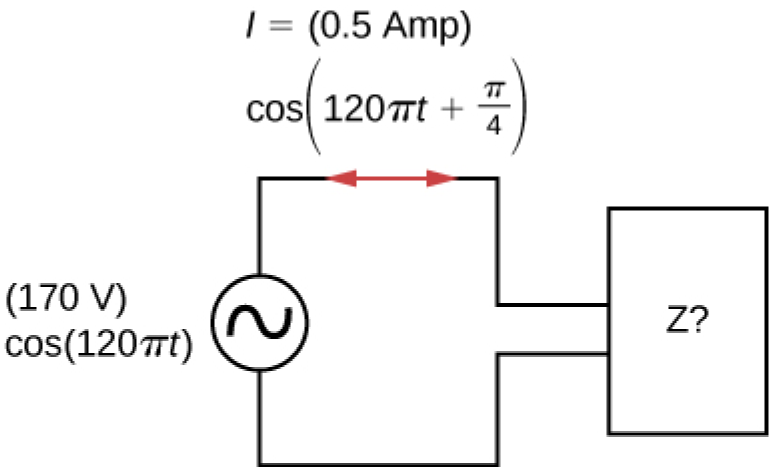
Problem:

Suppose a coil has a self-inductance of 20.0 H and a resistance of $200\,\Omega$. What (a) capacitance and (b) resistance must be connected in series with the coil to produce a circuit that has a resonant frequency of 100 Hz and a *Q* of 10?

Exercise:

Problem:

An ac generator is connected to a device whose internal circuits are not known. We only know current and voltage outside the device, as shown below. Based on the information given, what can you infer about the electrical nature of the device and its power usage?



Solution:

The reactance of the capacitor is larger than the reactance of the inductor because the current leads the voltage. The power usage is 30 W.

Glossary

bandwidth
: range of angular frequencies over which the average power is greater than one-half the maximum value of the average power

quality factor
: dimensionless quantity that describes the sharpness of the peak of the bandwidth; a high quality factor is a sharp or narrow resonance peak

resonant frequency

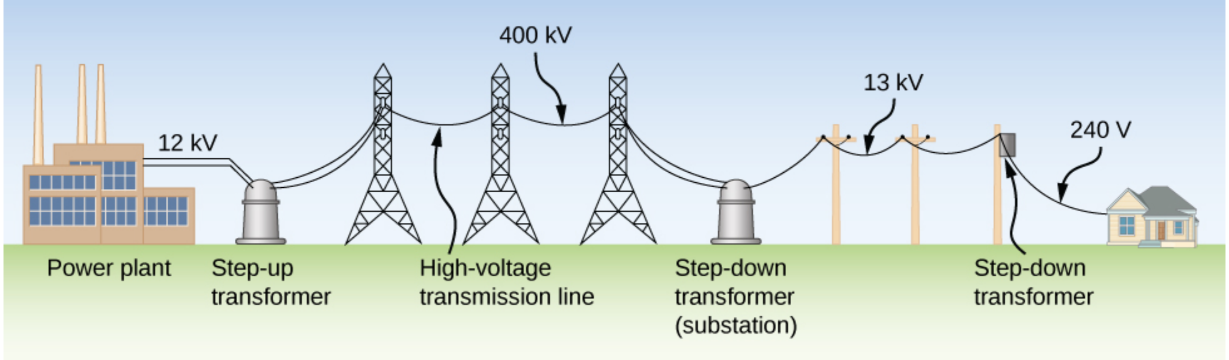
frequency at which the amplitude of the current is a maximum and the circuit would oscillate if not driven by a voltage source

Transformers

By the end of the section, you will be able to:

- Explain why power plants transmit electricity at high voltages and low currents and how they do this
- Develop relationships among current, voltage, and the number of windings in step-up and step-down transformers

Although ac electric power is produced at relatively low voltages, it is sent through transmission lines at very high voltages (as high as 500 kV). The same power can be transmitted at different voltages because power is the product $I_{\text{rms}}V_{\text{rms}}$. (For simplicity, we ignore the phase factor $\cos\phi$.) A particular power requirement can therefore be met with a low voltage and a high current or with a high voltage and a low current. The advantage of the high-voltage/low-current choice is that it results in lower $I_{\text{rms}}^2 R$ ohmic losses in the transmission lines, which can be significant in lines that are many kilometers long ([link]).

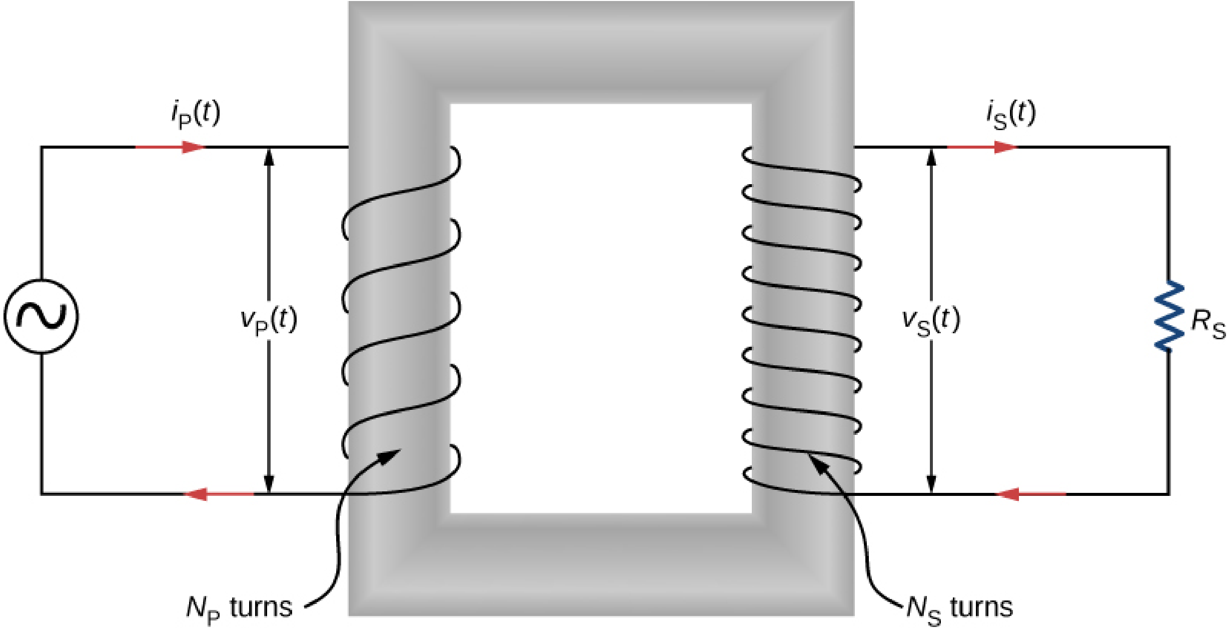


The rms voltage from a power plant eventually needs to be stepped down from 12 kV to 240 V so that it can be safely introduced into a home. A high-voltage transmission line allows a low current to be transmitted via a substation over long distances.

Typically, the alternating emfs produced at power plants are “stepped up” to very high voltages before being transmitted through power lines; then, they must be “stepped down” to relatively safe values (110 or 220 V rms) before they are introduced into homes. The device that transforms voltages from one value to another using induction is the **transformer** ([link]).

Transformers are used to step down the high voltages in transmission lines to the 110 to 220 V used in homes. (credit: modification of work by “Fortyseven”/Flickr)

As [link] illustrates, a transformer basically consists of two separated coils, or windings, wrapped around a soft iron core. The primary winding has N_{P} loops, or turns, and is connected to an alternating voltage $v_{\text{P}}(t)$. The secondary winding has N_{S} turns and is connected to a load resistor R_{S}. We assume the ideal case for which all magnetic field lines are confined to the core so that the same magnetic flux permeates each turn of both the primary and the secondary windings. We also neglect energy losses to magnetic hysteresis, to ohmic heating in the windings, and to ohmic heating of the induced eddy currents in the core. A good transformer can have losses as low as 1% of the transmitted power, so this is not a bad assumption.



A step-up transformer (more turns in the secondary winding than in the primary winding). The two windings are wrapped around a soft iron core.

To analyze the transformer circuit, we first consider the primary winding. The input voltage $v_P(t)$ is equal to the potential difference induced across the primary winding. From Faraday's law, the induced potential difference is $-N_P\,(d\Phi/dt)$, where Φ is the flux through one turn of the primary winding. Thus,

Equation:

$$v_P\,(t) = -N_P\frac{d\Phi}{dt}.$$

Similarly, the output voltage $v_S(t)$ delivered to the load resistor must equal the potential difference induced across the secondary winding. Since the transformer is ideal, the flux through every turn of the secondary winding is also Φ, and

Equation:

$$v_{\mathrm{S}}(t) = -N_{\mathrm{S}}\frac{d\Phi}{dt}.$$

Combining the last two equations, we have
Equation:

$$v_{\mathrm{S}}\left(t\right) = \frac{N_{\mathrm{S}}}{N_{\mathrm{P}}}v_{\mathrm{P}}\left(t\right).$$

Hence, with appropriate values for N_{S} and N_{P}, the input voltage $v_{\mathrm{P}}(t)$ may be "stepped up" $\left(N_{\mathrm{S}} > N_{\mathrm{P}}\right)$ or "stepped down" $(N_{\mathrm{S}} < N_{\mathrm{P}})$ to $v_{\mathrm{S}}(t)$, the output voltage. This is often abbreviated as the **transformer equation**,

Note:
Equation:

$$\frac{V_{\mathrm{S}}}{V_{\mathrm{P}}} = \frac{N_{\mathrm{S}}}{N_{\mathrm{P}}},$$

which shows that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of turns in their windings. For a **step-up transformer**, which increases voltage and decreases current, this ratio is greater than one; for a **step-down transformer**, which decreases voltage and increases current, this ratio is less than one.

From the law of energy conservation, the power introduced at any instant by $v_{\mathrm{P}}(t)$ to the primary winding must be equal to the power dissipated in the resistor of the secondary circuit; thus,
Equation:

$$i_{\mathrm{P}}(t)v_{\mathrm{P}}(t) = i_{\mathrm{S}}(t)v_{\mathrm{S}}(t).$$

When combined with [link], this gives

Equation:

$$i_{\mathrm{S}}(t) = \frac{N_{\mathrm{P}}}{N_{\mathrm{S}}} i_{\mathrm{P}}(t).$$

If the voltage is stepped up, the current is stepped down, and vice versa.

Finally, we can use $i_{\mathrm{S}}(t) = v_{\mathrm{S}}(t)/R_{\mathrm{S}}$, along with [link] and [link], to obtain

Equation:

$$v_{\mathrm{P}}(t) = i_{\mathrm{P}} \left[\left(\frac{N_{\mathrm{P}}}{N_{\mathrm{S}}} \right)^2 R_{\mathrm{S}} \right],$$

which tells us that the input voltage $v_{\mathrm{P}}(t)$ “sees” not a resistance R_{S} but rather a resistance

Equation:

$$R_{\mathrm{P}} = \left(\frac{N_{\mathrm{P}}}{N_{\mathrm{S}}} \right)^2 R_{\mathrm{S}}.$$

Our analysis has been based on instantaneous values of voltage and current. However, the resulting equations are not limited to instantaneous values; they hold also for maximum and rms values.

Example:

A Step-Down Transformer

A transformer on a utility pole steps the rms voltage down from 12 kV to 240 V. (a) What is the ratio of the number of secondary turns to the number of primary turns? (b) If the input current to the transformer is 2.0 A, what is the output current? (c) Determine the power loss in the transmission line if the total resistance of the transmission line is 200 Ω. (d) What would the power loss have been if the transmission line was at 240 V the entire length of the line, rather than providing voltage at 12 kV? What does this say about transmission lines?

Strategy

The number of turns related to the voltages is found from [link]. The output current is calculated using [link].

Solution

a. Using [link] with rms values V_P and V_S, we have

Equation:

$$\frac{N_\text{S}}{N_\text{P}} = \frac{240\text{ V}}{12 \times 10^3\text{ V}} = \frac{1}{50},$$

so the primary winding has 50 times the number of turns in the secondary winding.

b. From [link], the output rms current I_S is found using the transformer equation with current

Note:

Equation:

$$I_\text{S} = \frac{N_\text{P}}{N_\text{S}} I_\text{P}$$

such that

Equation:

$$I_\text{S} = \frac{N_\text{P}}{N_\text{S}} I_\text{P} = (50)\,(2.0\text{ A}) = 100\text{ A}.$$

c. The power loss in the transmission line is calculated to be

Equation:

$$P_\text{loss} = I_\text{P}^2 R = (2.0\text{ A})^2\,(200\,\Omega) = 800\text{ W}.$$

d. If there were no transformer, the power would have to be sent at 240 V to work for these houses, and the power loss would be

Equation:

$$P_{\text{loss}} = I_S^2 R = (100\text{ A})^2\,(200\,\Omega) = 2\,\times\,10^6\text{ W}.$$

Therefore, when power needs to be transmitted, we want to avoid power loss. Thus, lines are sent with high voltages and low currents and adjusted with a transformer before power is sent into homes.

Significance
This application of a step-down transformer allows a home that uses 240-V outlets to have 100 A available to draw upon. This can power many devices in the home.

Note:
Exercise:

Problem:

Check Your Understanding A transformer steps the line voltage down from 110 to 9.0 V so that a current of 0.50 A can be delivered to a doorbell. (a) What is the ratio of the number of turns in the primary and secondary windings? (b) What is the current in the primary winding? (c) What is the resistance seen by the 110-V source?

Solution:

a. 12:1; b. 0.042 A; c. $2.6\,\times\,10^3\,\Omega$

Summary

- Power plants transmit high voltages at low currents to achieve lower ohmic losses in their many kilometers of transmission lines.
- Transformers use induction to transform voltages from one value to another.
- For a transformer, the voltages across the primary and secondary coils, or windings, are related by the transformer equation.

- The currents in the primary and secondary windings are related by the number of primary and secondary loops, or turns, in the windings of the transformer.
- A step-up transformer increases voltage and decreases current, whereas a step-down transformer decreases voltage and increases current.

Key Equations

AC voltage	$v = V_0 \sin \omega t$
AC current	$i = I_0 \sin \omega t$
capacitive reactance	$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C$
rms voltage	$V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$
rms current	$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$
inductive reactance	$\frac{V_0}{I_0} = \omega L = X_L$
Phase angle of an RLC series circuit	$\phi = \tan^{-1} \frac{X_L - X_C}{R}$
AC version of Ohm’s law	$I_0 = \frac{V_0}{Z}$
Impedance of an RLC series circuit	$Z = \sqrt{R^2 + (X_L - X_C)^2}$
Average power associated with a circuit element	$P_{\text{ave}} = \frac{1}{2} I_0 V_0 \cos \phi$
Average power dissipated by a resistor	$P_{\text{ave}} = \frac{1}{2} I_0 V_0 = I_{\text{rms}} V_{\text{rms}} = I_{\text{rms}}^2 R$

Resonant angular frequency of a circuit	$\omega_0 = \sqrt{\frac{1}{LC}}$
Quality factor of a circuit	$Q = \frac{\omega_0}{\Delta\omega}$
Quality factor of a circuit in terms of the circuit parameters	$Q = \frac{\omega_0 L}{R}$
Transformer equation with voltage	$\frac{V_\text{S}}{V_\text{P}} = \frac{N_\text{S}}{N_\text{P}}$
Transformer equation with current	$I_\text{S} = \frac{N_\text{P}}{N_\text{S}} I_\text{P}$

Conceptual Questions

Exercise:

Problem:

Why do transmission lines operate at very high voltages while household circuits operate at fairly small voltages?

Solution:

There is less thermal loss if the transmission lines operate at low currents and high voltages.

Exercise:

Problem:

How can you distinguish the primary winding from the secondary winding in a step-up transformer?

Exercise:

Problem:

Battery packs in some electronic devices are charged using an adapter connected to a wall socket. Speculate as to the purpose of the adapter.

Solution:

The adapter has a step-down transformer to have a lower voltage and possibly higher current at which the device can operate.

Exercise:

Problem: Will a transformer work if the input is a dc voltage?

Exercise:

Problem:

Why are the primary and secondary coils of a transformer wrapped around the same closed loop of iron?

Solution:

so each loop can experience the same changing magnetic flux

Problems

Exercise:

Problem:

A step-up transformer is designed so that the output of its secondary winding is 2000 V (rms) when the primary winding is connected to a 110-V (rms) line voltage. (a) If there are 100 turns in the primary winding, how many turns are there in the secondary winding? (b) If a resistor connected across the secondary winding draws an rms current of 0.75 A, what is the current in the primary winding?

Exercise:

Problem:

A step-up transformer connected to a 110-V line is used to supply a hydrogen-gas discharge tube with 5.0 kV (rms). The tube dissipates 75 W of power. (a) What is the ratio of the number of turns in the secondary winding to the number of turns in the primary winding? (b) What are the rms currents in the primary and secondary windings? (c) What is the effective resistance seen by the 110-V source?

Solution:

a. 45:1; b. 0.68 A, 0.015 A; c. 160 Ω

Exercise:

Problem:

An ac source of emf delivers 5.0 mW of power at an rms current of 2.0 mA when it is connected to the primary coil of a transformer. The rms voltage across the secondary coil is 20 V. (a) What are the voltage across the primary coil and the current through the secondary coil? (b) What is the ratio of secondary to primary turns for the transformer?

Exercise:

Problem:

A transformer is used to step down 110 V from a wall socket to 9.0 V for a radio. (a) If the primary winding has 500 turns, how many turns does the secondary winding have? (b) If the radio operates at a current of 500 mA, what is the current through the primary winding?

Solution:

a. 41 turns; b. 40.9 mA

Exercise:

Problem:

A transformer is used to supply a 12-V model train with power from a 110-V wall plug. The train operates at 50 W of power. (a) What is the rms current in the secondary coil of the transformer? (b) What is the rms current in the primary coil? (c) What is the ratio of the number of primary to secondary turns? (d) What is the resistance of the train? (e) What is the resistance seen by the 110-V source?

Additional Problems

Exercise:

Problem:

The emf of an ac source is given by $v(t) = V_0 \sin \omega t$, where $V_0 = 100 \text{ V}$ and $\omega = 200\pi \text{ rad/s}$. Find an expression that represents the output current of the source if it is connected across (a) a $20\text{-}\mu\text{F}$ capacitor, (b) a 20-mH inductor, and (c) a $50\text{-}\Omega$ resistor.

Solution:

a. $i(t) = (1.26\text{A}) \sin (200\pi t + \pi/2)$; b. $i(t) = (7.96\text{A}) \sin (200\pi t - \pi/2)$; c. $i(t) = (2\text{A}) \sin (200\pi t)$

Exercise:

Problem:

A 700-pF capacitor is connected across an ac source with a voltage amplitude of 160 V and a frequency of 20 kHz. (a) Determine the capacitive reactance of the capacitor and the amplitude of the output current of the source. (b) If the frequency is changed to 60 Hz while keeping the voltage amplitude at 160 V, what are the capacitive reactance and the current amplitude?

Exercise:

Problem:

A 20-mH inductor is connected across an AC source with a variable frequency and a constant-voltage amplitude of 9.0 V. (a) Determine the reactance of the circuit and the maximum current through the inductor when the frequency is set at 20 kHz. (b) Do the same calculations for a frequency of 60 Hz.

Solution:

a. $2.5 \times 10^3 \Omega, 3.6 \times 10^{-3} \text{ A}$; b. $7.5 \, \Omega, 1.2\text{A}$

Exercise:

Problem:

A 30-μF capacitor is connected across a 60-Hz ac source whose voltage amplitude is 50 V. (a) What is the maximum charge on the capacitor? (b) What is the maximum current into the capacitor? (c) What is the phase relationship between the capacitor charge and the current in the circuit?

Exercise:

Problem:

A 7.0-mH inductor is connected across a 60-Hz ac source whose voltage amplitude is 50 V. (a) What is the maximum current through the inductor? (b) What is the phase relationship between the current through and the potential difference across the inductor?

Solution:

a. 19 A; b. inductor leads by 90°

Exercise:

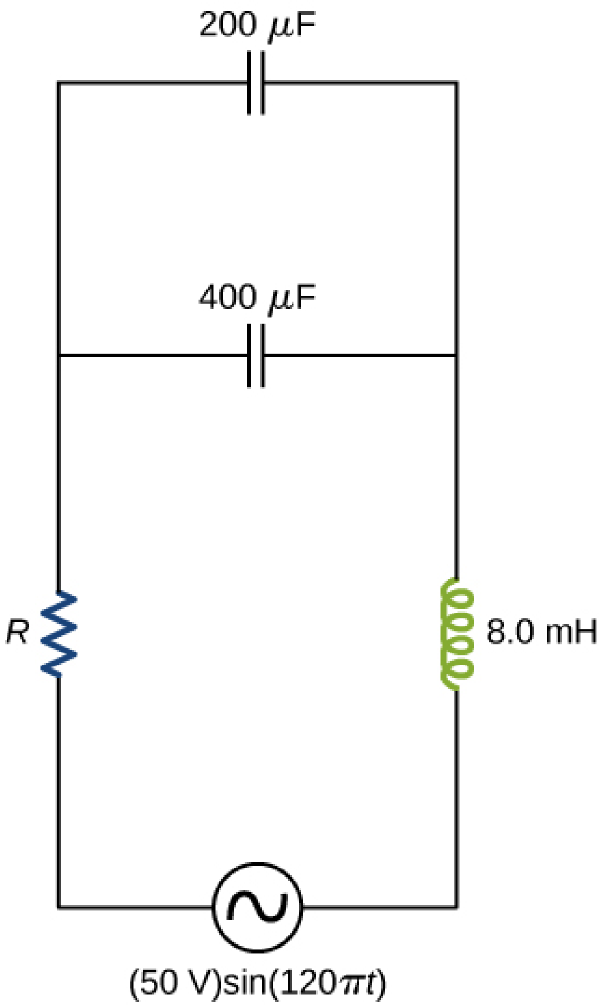
Problem:

What is the impedance of an *RLC* series circuit at the resonant frequency?

Exercise:

Problem:

What is the resistance *R* in the circuit shown below if the amplitude of the ac through the inductor is 4.24 A?



Solution:

$11.7\,\Omega$

Exercise:

Problem:

An ac source of voltage amplitude 100 V and frequency 1.0 kHz drives an *RLC* series circuit with $R = 20\,\Omega$, $L = 4.0\,\text{mH}$, and $C = 50\mu\text{F}$. (a) Determine the rms current through the circuit. (b) What are the rms voltages across the three elements? (c) What is the phase angle between the emf and the current? (d) What is the power output of the source? (e) What is the power dissipated in the resistor?

Exercise:

Problem:

In an *RLC* series circuit, $R = 200\,\Omega$, $L = 1.0\,\text{H}$, $C = 50\mu\text{F}$, $V_0 = 120\,\text{V}$, and $f = 50\,\text{Hz}$. What is the power output of the source?

Solution:

14 W

Exercise:

Problem:

A power plant generator produces 100 A at 15 kV (rms). A transformer is used to step up the transmission line voltage to 150 kV (rms). (a) What is rms current in the transmission line? (b) If the resistance per unit length of the line is $8.6\ \times\ 10^{-8}\,\Omega/\text{m}$, what is the power loss per meter in the line? (c) What would the power loss per meter be if the line voltage were 15 kV (rms)?

Exercise:

Problem:

Consider a power plant located 25 km outside a town delivering 50 MW of power to the town. The transmission lines are made of aluminum cables with a $7\,\text{cm}^2$ cross-sectional area. Find the loss of power in the transmission lines if it is transmitted at (a) 200 kV (rms) and (b) 120 V (rms).

Solution:

a. $5.9\ \times\ 10^4$ W; b. $1.64\ \times\ 10^{11}$ W

Exercise:

Problem:

Neon signs require 12-kV for their operation. A transformer is to be used to change the voltage from 220-V (rms) ac to 12-kV (rms) ac. What must the ratio be of turns in the secondary winding to the turns in the primary winding? (b) What is the maximum rms current the neon lamps can draw if the fuse in the primary winding goes off at 0.5 A? (c) How much power is used by the neon sign when it is drawing the maximum current allowed by the fuse in the primary winding?

Challenge Problems

Exercise:

Problem:

The 335-kV ac electricity from a power transmission line is fed into the primary winding of a transformer. The ratio of the number of turns in the secondary winding to the number in the primary winding is $N_{\text{s}}/N_{\text{p}}=1000$. (a) What voltage is induced in the secondary winding? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

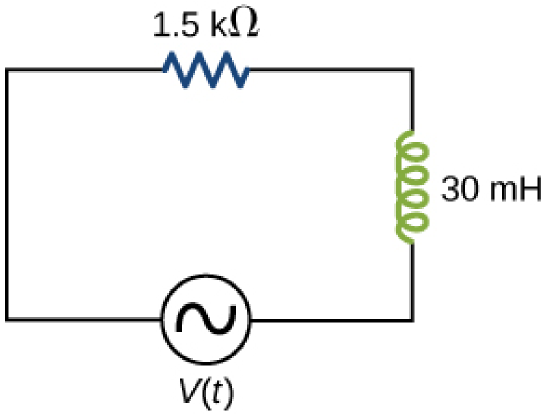
Solution:

a. 335 MV; b. the result is way too high, well beyond the breakdown voltage of air over reasonable distances; c. the input voltage is too high

Exercise:

Problem:

A 1.5-kΩ resistor and 30-mH inductor are connected in series, as shown below, across a 120-V (rms) ac power source oscillating at 60-Hz frequency. (a) Find the current in the circuit. (b) Find the voltage drops across the resistor and inductor. (c) Find the impedance of the circuit. (d) Find the power dissipated in the resistor. (e) Find the power dissipated in the inductor. (f) Find the power produced by the source.



Exercise:

Problem:

A 20-Ω resistor, 50-μF capacitor, and 30-mH inductor are connected in series with an ac source of amplitude 10 V and frequency 125 Hz. (a) What is the impedance of the circuit? (b) What is the amplitude of the current in the circuit? (c) What is the phase constant of the current? Is it leading or lagging the source voltage? (d) Write voltage drops across the resistor, capacitor, and inductor and the source voltage as a function of time. (e) What is the power factor of the circuit? (f) How much energy is used by the resistor in 2.5 s?

Solution:

a. 20 Ω; b. 0.5 A; c. 5.4°, lagging;
d.
$V_R = (9.96\text{ V})\cos(250\pi t + 5.4^\circ), V_C = (12.7\text{ V})\cos(250\pi t + 5.4^\circ - 90^\circ),$
$V_L = (11.8\text{ V})\cos(250\pi t + 5.4^\circ + 90^\circ), V_{\text{source}} = (10.0\text{ V})\cos(250\pi t);$
e. 0.995; f. 6.25 J

Exercise:

Problem:

A 200-Ω resistor, 150-μF capacitor, and 2.5-H inductor are connected in series with an ac source of amplitude 10 V and variable angular frequency ω. (a) What is the value of the resonance frequency ω_R? (b) What is the amplitude of the current if $\omega = \omega_R$? (c) What is the phase constant of the current when $\omega = \omega_R$? Is it leading or lagging the source voltage, or is it in phase? (d) Write an equation for the voltage drop across the resistor as a function of time when $\omega = \omega_R$. (e) What is the power factor of the circuit when $\omega = \omega_R$? (f) How much energy is used up by the resistor in 2.5 s when $\omega = \omega_R$?

Exercise:

Problem:

Find the reactances of the following capacitors and inductors in ac circuits with the given frequencies in each case: (a) 2-mH inductor with a frequency 60-Hz of the ac circuit; (b) 2-mH inductor with a frequency 600-Hz of the ac circuit; (c) 20-mH inductor with a frequency 6-Hz of the ac circuit; (d) 20-mH inductor with a frequency 60-Hz of the ac circuit; (e) 2-mF capacitor with a frequency 60-Hz of the ac circuit; and (f) 2-mF capacitor with a frequency 600-Hz of the AC circuit.

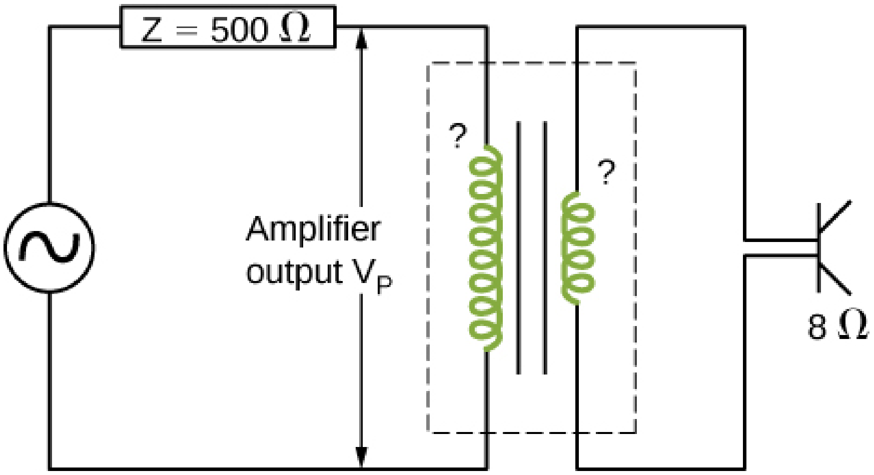
Solution:

a. $0.75\ \Omega$; b. $7.5\ \Omega$; c. $0.75\ \Omega$; d. $7.5\ \Omega$; e. $1.3\ \Omega$; f. $0.13\ \Omega$

Exercise:

Problem:

An output impedance of an audio amplifier has an impedance of $500\ \Omega$ and has a mismatch with a low-impedance 8-Ω loudspeaker. You are asked to insert an appropriate transformer to match the impedances. What turns ratio will you use, and why? Use the simplified circuit shown below.



Exercise:

Problem:

Show that the SI unit for capacitive reactance is the ohm. Show that the SI unit for inductive reactance is also the ohm.

Solution:

The units as written for inductive reactance [link] are $\frac{\text{rad}}{\text{s}}\text{H}$. Radians can be ignored in unit analysis. The Henry can be defined as $\text{H} = \frac{\text{V·s}}{\text{A}} = \Omega \cdot \text{s}$. Combining these together results in a unit of Ω for reactance.

Exercise:

Problem:

A coil with a self-inductance of 16 mH and a resistance of $6.0\ \Omega$ is connected to an ac source whose frequency can be varied. At what frequency will the voltage across the coil lead the current through the coil by 45°?

Exercise:

Problem:

An *RLC* series circuit consists of a $50\text{-}\Omega$ resistor, a $200\text{-}\mu\text{F}$ capacitor, and a 120-mH inductor whose coil has a resistance of $20\ \Omega$. The source for the circuit has an rms emf of 240 V at a frequency of 60 Hz. Calculate the rms voltages across the (a) resistor, (b) capacitor, and (c) inductor.

Solution:

a. 156 V; b. 42 V; c. 154 V

Exercise:

Problem:

An *RLC* series circuit consists of a $10\text{-}\Omega$ resistor, an $8.0\text{-}\mu\text{F}$ capacitor, and a 50-mH inductor. A 110-V (rms) source of variable frequency is connected across the combination. What is the power output of the source when its frequency is set to one-half the resonant frequency of the circuit?

Exercise:

Problem:

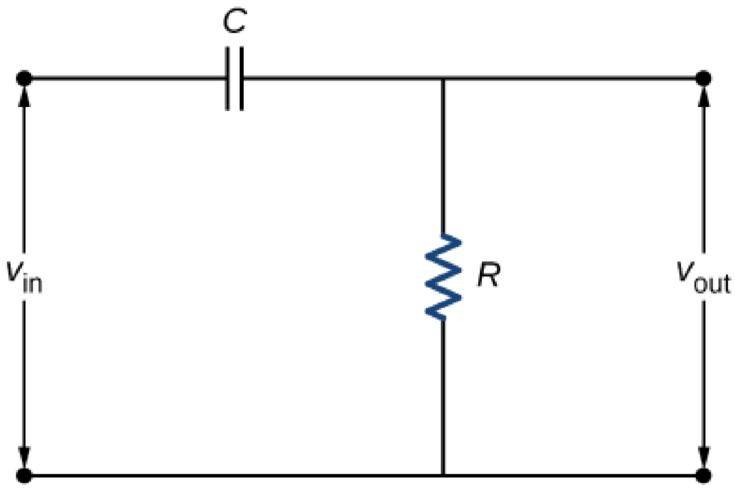
Shown below are two circuits that act as crude high-pass filters. The input voltage to the circuits is v_{in}, and the output voltage is v_{out}. (a) Show that for the capacitor circuit,

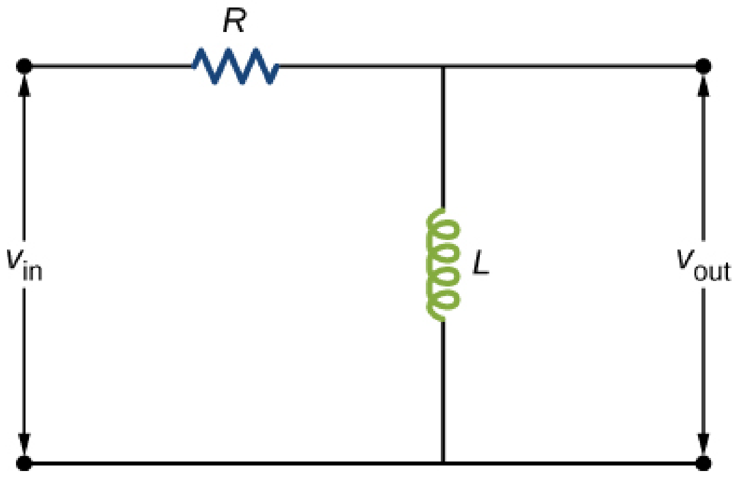
$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{1}{\sqrt{1+1/\omega^2 R^2 C^2}},$$

and for the inductor circuit,

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{\omega L}{\sqrt{R^2+\omega^2 L^2}}.$$

(b) Show that for high frequencies, $v_{\text{out}} \approx v_{\text{in}}$, but for low frequencies, $v_{\text{out}} \approx 0$.





Solution:

a. $\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{1}{\sqrt{1+1/\omega^2R^2C^2}}$ and $\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{\omega L}{\sqrt{R^2+\omega^2L^2}}$; b. $v_{\text{out}} \approx v_{\text{in}}$ and $v_{\text{out}} \approx 0$

Exercise:

Problem:

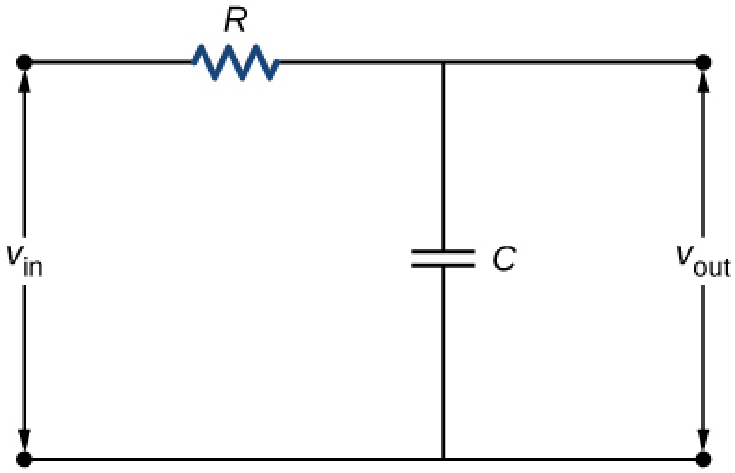
The two circuits shown below act as crude low-pass filters. The input voltage to the circuits is v_{in}, and the output voltage is v_{out}. (a) Show that for the capacitor circuit,

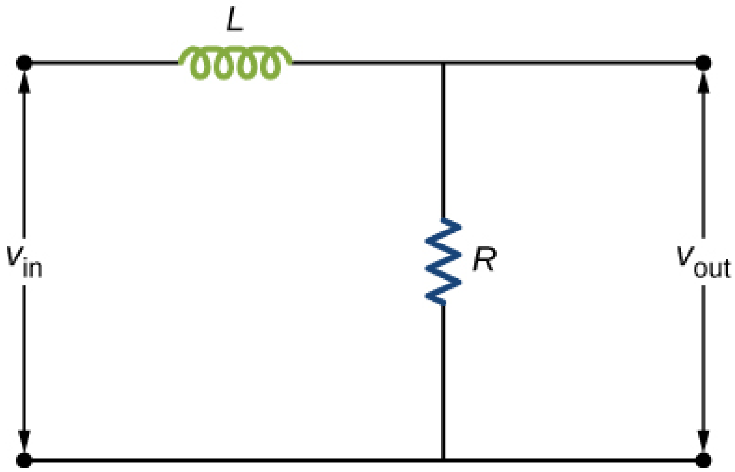
$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{1}{\sqrt{1+\omega^2R^2C^2}},$$

and for the inductor circuit,

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{R}{\sqrt{R^2+\omega^2 L^2}}.$$

(b) Show that for low frequencies, $v_{\text{out}} \approx v_{\text{in}}$, but for high frequencies, $v_{\text{out}} \approx 0$.


R
v_{in}
C
v_{out}


L
v_{in}
R
v_{out}

Glossary

step-down transformer

transformer that decreases voltage and increases current

step-up transformer

transformer that increases voltage and decreases current

transformer
device that transforms voltages from one value to another using induction

transformer equation
equation showing that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of turns in their windings

Introduction class="introduction"

The pressure from sunlight predicted by Maxwell’s equations helped produce the tail of Comet McNaught. (credit: modification of work by Sebastian Deiries—ESO)

Our view of objects in the sky at night, the warm radiance of sunshine, the sting of sunburn, our cell phone conversations, and the X-rays revealing a broken bone—all are brought to us by electromagnetic waves. It would be hard to overstate the practical importance of electromagnetic waves, through their role in vision, through countless technological applications, and through their ability to transport the energy from the Sun through space to sustain life and almost all of its activities on Earth.

Theory predicted the general phenomenon of electromagnetic waves before anyone realized that light is a form of an electromagnetic wave. In the mid-nineteenth century, James Clerk Maxwell formulated a single theory combining all the electric and magnetic effects known at that time. Maxwell’s equations, summarizing this theory, predicted the existence of electromagnetic waves that travel at the speed of light. His theory also predicted how these waves behave, and how they carry both energy and momentum. The tails of comets, such as Comet McNaught in [link], provide a spectacular example. Energy carried by light from the Sun warms the comet to release dust and gas. The momentum carried by the light exerts a weak force that shapes the dust into a tail of the kind seen here. The flux of particles emitted by the Sun, called the solar wind, typically produces an additional, second tail, as described in detail in this chapter.

In this chapter, we explain Maxwell’s theory and show how it leads to his prediction of electromagnetic waves. We use his theory to examine what electromagnetic waves are, how they are produced, and how they transport energy and momentum. We conclude by summarizing some of the many practical applications of electromagnetic waves.

Maxwell’s Equations and Electromagnetic Waves

By the end of this section, you will be able to:

- Explain Maxwell’s correction of Ampère’s law by including the displacement current
- State and apply Maxwell’s equations in integral form
- Describe how the symmetry between changing electric and changing magnetic fields explains Maxwell’s prediction of electromagnetic waves
- Describe how Hertz confirmed Maxwell’s prediction of electromagnetic waves

James Clerk Maxwell (1831–1879) was one of the major contributors to physics in the nineteenth century ([link]). Although he died young, he made major contributions to the development of the kinetic theory of gases, to the understanding of color vision, and to the nature of Saturn’s rings. He is probably best known for having combined existing knowledge of the laws of electricity and of magnetism with insights of his own into a complete overarching electromagnetic theory, represented by **Maxwell’s equations**.

James Clerk Maxwell, a nineteenth-century physicist, developed a theory that explained the relationship between electricity and magnetism, and correctly predicted that visible light consists of electromagnetic waves.

Maxwell’s Correction to the Laws of Electricity and Magnetism

The four basic laws of electricity and magnetism had been discovered experimentally through the work of physicists such as Oersted, Coulomb, Gauss, and Faraday. Maxwell discovered logical inconsistencies in these earlier results and identified the incompleteness of Ampère’s law as their cause.

Recall that according to Ampère’s law, the integral of the magnetic field around a closed loop C is proportional to the current I passing through any surface whose boundary is loop C itself:
Equation:

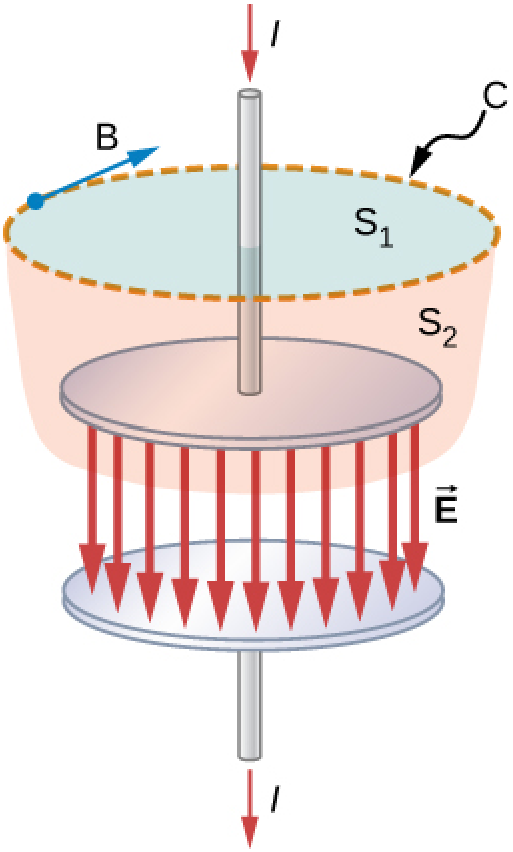
$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I.$$

There are infinitely many surfaces that can be attached to any loop, and Ampère’s law stated in [link] is independent of the choice of surface.

Consider the set-up in [link]. A source of emf is abruptly connected across a parallel-plate capacitor so that a time-dependent current I develops in the wire. Suppose we apply Ampère’s law to loop C shown at a time before the capacitor is fully charged, so that $I \neq 0$. Surface S_1 gives a nonzero value for the enclosed current I, whereas surface S_2 gives zero for the enclosed current because no current passes through it:
Equation:

$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \begin{cases} \mu_0 I & \text{if surface } S_1 \text{ is used} \\ 0 & \text{if surface } S_2 \text{ is used} \end{cases}.$$

Clearly, Ampère’s law in its usual form does not work here. This may not be surprising, because Ampère’s law as applied in earlier chapters required a steady current, whereas the current in this experiment is changing with time and is not steady at all.



The currents through surface S_1 and surface S_2 are unequal, despite having the same boundary loop C.

How can Ampère's law be modified so that it works in all situations? Maxwell suggested including an additional contribution, called the displacement current I_d, to the real current I,

Equation:

$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 \left(I + I_\text{d} \right)$$

where the displacement current is defined to be

Note:
Equation:

$$I_{\text{d}} = \varepsilon_0 \frac{d\Phi_{\text{E}}}{dt}.$$

Here ε_0 is the permittivity of free space and Φ_{E} is the electric flux, defined as

Equation:

$$\Phi_{\text{E}} = \iint_{\text{Surface } S} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}.$$

The **displacement current** is analogous to a real current in Ampère’s law, entering into Ampère’s law in the same way. It is produced, however, by a changing electric field. It accounts for a changing electric field producing a magnetic field, just as a real current does, but the displacement current can produce a magnetic field even where no real current is present. When this extra term is included, the modified Ampère’s law equation becomes

Equation:

$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I + \varepsilon_0 \mu_0 \frac{d\Phi_{\text{E}}}{dt}$$

and is independent of the surface *S* through which the current *I* is measured.

We can now examine this modified version of Ampère’s law to confirm that it holds independent of whether the surface S_1 or the surface S_2 in [link] is chosen. The electric field $\vec{\mathbf{E}}$ corresponding to the flux Φ_E in [link] is between the capacitor plates. Therefore, the $\vec{\mathbf{E}}$ field and the displacement current through the surface S_1 are both zero, and [link] takes the form

Equation:

$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I.$$

We must now show that for surface S_2, through which no actual current flows, the displacement current leads to the same value $\mu_0 I$ for the right side of the Ampère's law equation. For surface S_2, the equation becomes

Equation:

$$\oint_C \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 \frac{d}{dt}\left[\varepsilon_0 \iint_{\text{Surface } S_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}\right].$$

Gauss's law for electric charge requires a closed surface and cannot ordinarily be applied to a surface like S_1 alone or S_2 alone. But the two surfaces S_1 and S_2 form a closed surface in [link] and can be used in Gauss's law. Because the electric field is zero on S_1, the flux contribution through S_1 is zero. This gives us

Equation:

$$\begin{aligned}\oiint_{\text{Surface } S_1+S_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} &= \iint_{\text{Surface } S_1} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} + \iint_{\text{Surface } S_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} \\ &= 0 + \iint_{\text{Surface } S_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} \\ &= \iint_{\text{Surface } S_2} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}.\end{aligned}$$

Therefore, we can replace the integral over S_2 in [link] with the closed Gaussian surface $S_1 + S_2$ and apply Gauss's law to obtain

Equation:

$$\oint_{S_1} \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 \frac{dQ_{\text{in}}}{dt} = \mu_0 I.$$

Thus, the modified Ampère's law equation is the same using surface S_2, where the right-hand side results from the displacement current, as it is for the surface S_1, where the contribution comes from the actual flow of electric charge.

Example:

Displacement current in a charging capacitor

A parallel-plate capacitor with capacitance *C* whose plates have area *A* and separation distance *d* is connected to a resistor *R* and a battery of voltage *V*. The current starts to flow at $t = 0$. (a) Find the displacement current between the capacitor plates at time *t*. (b) From the properties of the capacitor, find the corresponding real current $I = \frac{dQ}{dt}$, and compare the answer to the expected current in the wires of the corresponding *RC* circuit.

Strategy

We can use the equations from the analysis of an *RC* circuit (Alternating-Current Circuits) plus Maxwell's version of Ampère's law.

Solution

a. The voltage between the plates at time *t* is given by
   Equation:

$$V_C = \frac{1}{C}Q\left(t\right) = V_0\left(1 - e^{-t/RC}\right).$$

   Let the *z*-axis point from the positive plate to the negative plate. Then the *z*-component of the electric field between the plates as a function of time *t* is
   Equation:

$$E_z\left(t\right) = \frac{V_0}{d}\left(1 - e^{-t/RC}\right).$$

   Therefore, the z-component of the displacement current I_d between

the plates is
Equation:

$$I_{\mathrm{d}}(t) = \varepsilon_0 A \frac{\partial E_z(t)}{\partial t} = \varepsilon_0 A \frac{V_0}{d} \times \frac{1}{RC} e^{-t/RC} = \frac{V_0}{R} e^{-t/RC},$$

where we have used $C = \varepsilon_0 \frac{A}{d}$ for the capacitance.

b. From the expression for V_C, the charge on the capacitor is
Equation:

$$Q(t) = CV_C = CV_0 \left(1 - e^{-t/RC}\right).$$

The current into the capacitor after the circuit is closed, is therefore
Equation:

$$I = \frac{dQ}{dt} = \frac{V_0}{R} e^{-t/RC}.$$

This current is the same as I_{d} found in (a).

Maxwell’s Equations

With the correction for the displacement current, Maxwell’s equations take the form

Note:
Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = \frac{Q_{\text{in}}}{\varepsilon_0} \qquad \left(\text{Gauss's law}\right)$$

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}} = 0 \qquad \left(\text{Gauss's law for magnetism}\right)$$

Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d\Phi_{\text{m}}}{dt} \qquad \left(\text{Faraday's law}\right)$$

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I + \varepsilon_0 \mu_0 \frac{d\Phi_{\text{E}}}{dt} \qquad \left(\text{Ampère-Maxwell law}\right).$$

Once the fields have been calculated using these four equations, the Lorentz force equation
Equation:

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}} + q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$$

gives the force that the fields exert on a particle with charge q moving with velocity $\vec{\mathbf{v}}$. The Lorentz force equation combines the force of the electric field and of the magnetic field on the moving charge. The magnetic and electric forces have been examined in earlier modules. These four Maxwell's equations are, respectively,

Note:
Maxwell's Equations

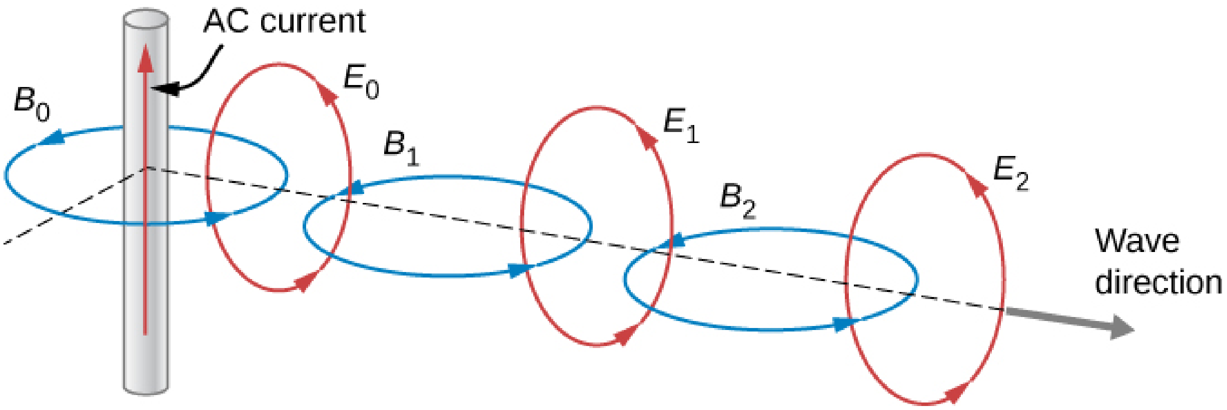
1. **Gauss's law**
The electric flux through any closed surface is equal to the electric charge Q_{in} enclosed by the surface. Gauss's law [[link]] describes the relation between an electric charge and the electric field it produces. This is often pictured in terms of electric field lines originating from positive charges and terminating on negative charges, and indicating the direction of the electric field at each point in space.
2. **Gauss's law for magnetism**
The magnetic field flux through any closed surface is zero [[link]]. This is equivalent to the statement that magnetic field lines are continuous, having no beginning or end. Any magnetic field line entering the region enclosed by the surface must also leave it. No magnetic monopoles, where magnetic field lines would terminate, are known to exist (see Magnetic Fields and Lines).
3. **Faraday's law**
A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. This third of Maxwell's equations, [link], is Faraday's law of induction and includes Lenz's law. The electric field from a changing magnetic field has field lines that form closed loops, without any beginning or end.
4. **Ampère-Maxwell law**
Magnetic fields are generated by moving charges or by changing electric fields. This fourth of Maxwell's equations, [link], encompasses Ampère's law and adds another source of magnetic fields, namely changing electric fields.

Maxwell's equations and the Lorentz force law together encompass all the laws of electricity and magnetism. The symmetry that Maxwell introduced into his mathematical framework may not be immediately apparent. Faraday's law describes how changing magnetic fields produce electric fields. The displacement current introduced by Maxwell results instead from a changing electric field and accounts for a changing electric field producing a magnetic field. The equations for the effects of both changing electric fields and changing magnetic fields differ in form only where the absence of magnetic monopoles leads to missing terms. This symmetry between the effects of changing magnetic and electric fields is essential in explaining the nature of electromagnetic waves.

Later application of Einstein’s theory of relativity to Maxwell’s complete and symmetric theory showed that electric and magnetic forces are not separate but are different manifestations of the same thing—the electromagnetic force. The electromagnetic force and weak nuclear force are similarly unified as the electroweak force. This unification of forces has been one motivation for attempts to unify all of the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces (see Particle Physics and Cosmology).

The Mechanism of Electromagnetic Wave Propagation

To see how the symmetry introduced by Maxwell accounts for the existence of combined electric and magnetic waves that propagate through space, imagine a time-varying magnetic field $\vec{\mathbf{B}}_0\,(t)$ produced by the high-frequency alternating current seen in [link]. We represent $\vec{\mathbf{B}}_0\,(t)$ in the diagram by one of its field lines. From Faraday’s law, the changing magnetic field through a surface induces a time-varying electric field $\vec{\mathbf{E}}_0\,(t)$ at the boundary of that surface. The displacement current source for the electric field, like the Faraday’s law source for the magnetic field, produces only closed loops of field lines, because of the mathematical symmetry involved in the equations for the induced electric and induced magnetic fields. A field line representation of $\vec{\mathbf{E}}_0\,(t)$ is shown. In turn, the changing electric field $\vec{\mathbf{E}}_0\,(t)$ creates a magnetic field $\vec{\mathbf{B}}_1\,(t)$ according to the modified Ampère’s law. This changing field induces $\vec{\mathbf{E}}_1\,(t)$, which induces $\vec{\mathbf{B}}_2\,(t)$, and so on. We then have a self-continuing process that leads to the creation of time-varying electric and magnetic fields in regions farther and farther away from O. This process may be visualized as the propagation of an electromagnetic wave through space.



How changing $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ fields propagate through space.

In the next section, we show in more precise mathematical terms how Maxwell’s equations lead to the prediction of electromagnetic waves that can travel through space without a material medium, implying a speed of electromagnetic waves equal to the speed of light.

Prior to Maxwell’s work, experiments had already indicated that light was a wave phenomenon, although the nature of the waves was yet unknown. In 1801, Thomas Young (1773–1829) showed that when a light beam was separated by two narrow slits and then recombined, a pattern made up of bright and dark fringes was formed on a screen. Young explained this behavior by assuming that light was composed of waves that added constructively at some points and destructively at others (see Interference). Subsequently, Jean Foucault (1819–1868), with measurements of the speed of light in various media, and Augustin Fresnel (1788–1827), with detailed experiments involving interference and diffraction of light, provided further conclusive evidence that light was a wave. So, light was known to be a wave, and Maxwell had predicted the existence of electromagnetic waves that traveled at the speed of light. The conclusion seemed inescapable: Light must be a form of electromagnetic radiation. But Maxwell’s theory showed that other wavelengths and frequencies than those of light were possible for electromagnetic waves. He showed that electromagnetic radiation with the same fundamental properties as visible light should exist at any frequency. It remained for others to test, and confirm, this prediction.

Note:
Exercise:

Problem:

Check Your Understanding When the emf across a capacitor is turned on and the capacitor is allowed to charge, when does the magnetic field induced by the displacement current have the greatest magnitude?

Solution:

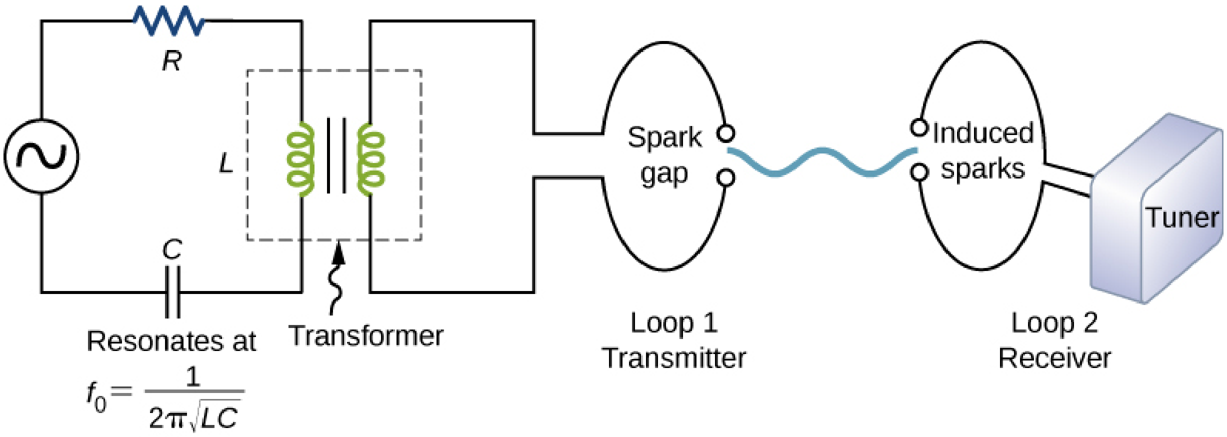
It is greatest immediately after the current is switched on. The displacement current and the magnetic field from it are proportional to the rate of change of electric field between the plates, which is greatest when the plates first begin to charge.

Hertz’s Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves but also verified that they travel at the speed of light.

Hertz used an alternating-current *RLC* (resistor-inductor-capacitor) circuit that resonates at a known frequency $f_0 = \frac{1}{2\pi\sqrt{LC}}$ and connected it to a loop of wire, as shown in [link]. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and helped generate electromagnetic waves.

Across the laboratory, Hertz placed another loop attached to another *RLC* circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could thus be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.


R
L
C
Resonates at
$f_0 = \frac{1}{2\pi\sqrt{LC}}$
Transformer
Spark gap
Induced sparks
Tuner
Loop 1
Transmitter
Loop 2
Receiver

The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, confirming their wave character. He was able to determine the wavelengths from the interference patterns, and knowing their frequencies, he could calculate the propagation speed using the equation $v = f\lambda$, where *v* is the speed of a wave, *f* is its frequency, and λ is its wavelength. Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz ($1\ \text{Hz} = 1\ \text{cycle/s}$), is named in his honor.

Note:
Exercise:

Problem:

Check Your Understanding Could a purely electric field propagate as a wave through a vacuum without a magnetic field? Justify your answer.

Solution:

No. The changing electric field according to the modified version of Ampère’s law would necessarily induce a changing magnetic field.

Summary

- Maxwell’s prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell’s equations.
- The four Maxwell’s equations together with the Lorentz force law encompass the major laws of electricity and magnetism. The first of these is Gauss’s law for electricity; the second is Gauss’s law for magnetism; the third is Faraday’s law of induction (including Lenz’s law); and the fourth is Ampère’s law in a symmetric formulation that adds another source of magnetism, namely changing electric fields.
- The symmetry introduced between electric and magnetic fields through Maxwell’s displacement current explains the mechanism of electromagnetic wave propagation, in which changing magnetic fields produce changing electric fields and vice versa.
- Although light was already known to be a wave, the nature of the wave was not understood before Maxwell. Maxwell’s equations also predicted electromagnetic waves with wavelengths and frequencies outside the range of light. These theoretical predictions were first confirmed experimentally by Heinrich Hertz.

Conceptual Questions

Exercise:

Problem:

Explain how the displacement current maintains the continuity of current in a circuit containing a capacitor.

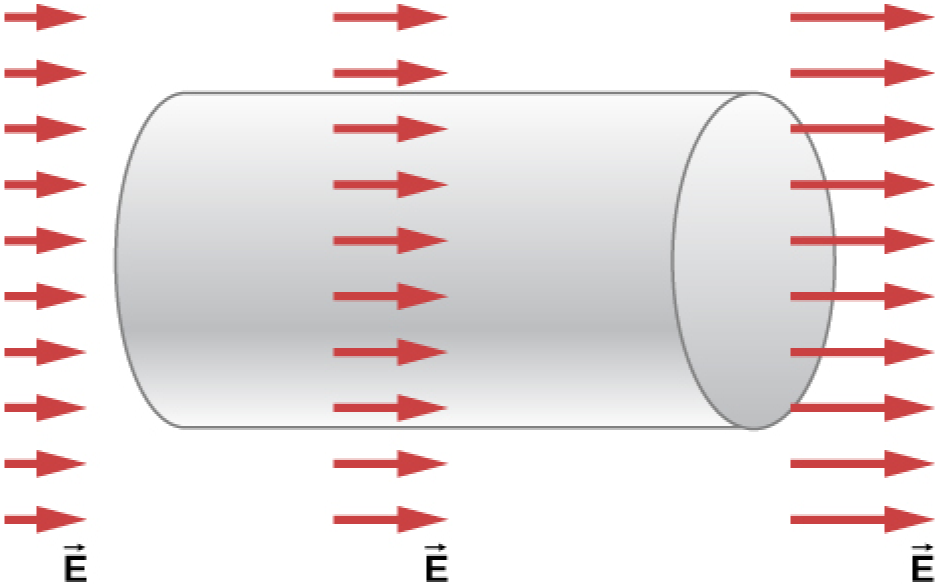
Solution:

The current into the capacitor to change the electric field between the plates is equal to the displacement current between the plates.

Exercise:

Problem:

Describe the field lines of the induced magnetic field along the edge of the imaginary horizontal cylinder shown below if the cylinder is in a spatially uniform electric field that is horizontal, pointing to the right, and increasing in magnitude.



Exercise:

Problem:

Why is it much easier to demonstrate in a student lab that a changing magnetic field induces an electric field than it is to demonstrate that a changing electric field produces a magnetic field?

Solution:

The first demonstration requires simply observing the current produced in a wire that experiences a changing magnetic field. The second demonstration requires moving electric charge from one location to another, and therefore involves electric currents that generate a changing electric field. The magnetic fields from these currents are not easily separated from the magnetic field that the displacement current produces.

Problems

Exercise:

Problem:

Show that the magnetic field at a distance *r* from the axis of two circular parallel plates, produced by placing charge $Q(t)$ on the plates is
$B_{\text{ind}} = \frac{\mu_0}{2\pi r}\frac{dQ(t)}{dt}$.

Solution:

$$\begin{aligned} B_{\text{ind}} &= \frac{\mu_0}{2\pi r} I_{\text{ind}} = \frac{\mu_0}{2\pi r}\varepsilon_0 \frac{\partial \Phi_E}{\partial t} = \frac{\mu_0}{2\pi r}\varepsilon_0 \left(A \frac{\partial E}{\partial t}\right) = \frac{\mu_0}{2\pi r}\varepsilon_0 A \left(\frac{1}{d}\ \frac{dV(t)}{dt}\right) \\ &= \frac{\mu_0}{2\pi r}\left[\frac{\varepsilon_0 A}{d}\right]\ \left[\frac{1}{C}\ \frac{dQ(t)}{dt}\right] = \frac{\mu_0}{2\pi r}\ \frac{dQ(t)}{dt} \qquad \text{because } C = \frac{\varepsilon_0 A}{d} \end{aligned}$$

Exercise:

Problem:

Express the displacement current in a capacitor in terms of the capacitance and the rate of change of the voltage across the capacitor.

Exercise:

Problem:

A potential difference $V(t) = V_0 \sin \omega t$ is maintained across a parallel-plate capacitor with capacitance C consisting of two circular parallel plates. A thin wire with resistance R connects the centers of the two plates, allowing charge to leak between plates while they are charging.
(a) Obtain expressions for the leakage current $I_{\text{res}}(t)$ in the thin wire. Use these results to obtain an expression for the current $I_{\text{real}}(t)$ in the wires connected to the capacitor.
(b) Find the displacement current in the space between the plates from the changing electric field between the plates.
(c) Compare $I_{\text{real}}(t)$ with the sum of the displacement current $I_{\text{d}}(t)$ and resistor current $I_{\text{res}}(t)$ between the plates, and explain why the relationship you observe would be expected.

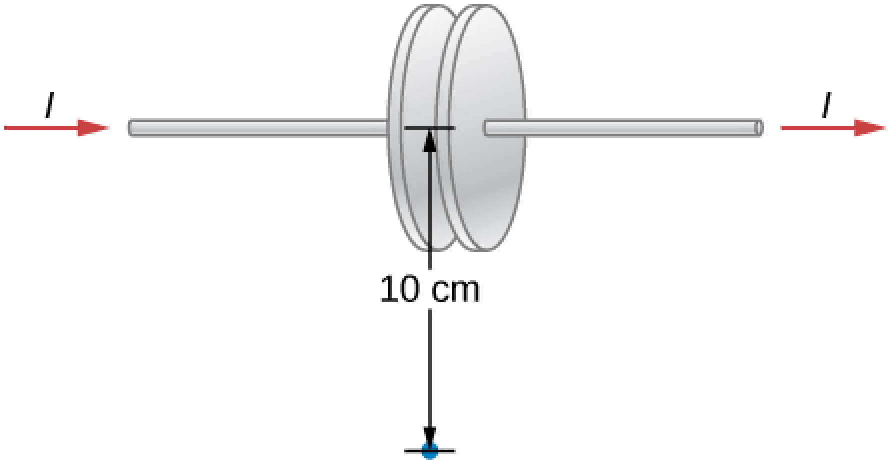
Solution:

a. $I_{\text{res}} = \frac{V_0 \sin \omega t}{R}$; b. $I_{\text{d}} = CV_0 \omega \cos \omega t$;
c.
$I_{\text{real}} = I_{\text{res}} + \frac{dQ}{dt} = \frac{V_0 \sin \omega t}{R} + CV_0 \frac{d}{dt} \sin \omega t = \frac{V_0 \sin \omega t}{R} + CV_0 \omega \cos \omega t$
; which is the sum of I_{res} and I_{real}, consistent with how the displacement current maintaining the continuity of current.

Exercise:

Problem:

Suppose the parallel-plate capacitor shown below is accumulating charge at a rate of 0.010 C/s. What is the induced magnetic field at a distance of 10 cm from the capacitator?



Exercise:

Problem:

The potential difference $V(t)$ between parallel plates shown above is instantaneously increasing at a rate of $10^7\ \text{V/s}$. What is the displacement current between the plates if the separation of the plates is 1.00 cm and they have an area of $0.200\ \text{m}^2$?

Solution:

$1.77\ \times\ 10^{-3}\ \text{A}$

Exercise:

Problem:

A parallel-plate capacitor has a plate area of $A = 0.250\ \text{m}^2$ and a separation of 0.0100 m. What must be must be the angular frequency ω for a voltage $V\ (t) = V_0\ \sin \omega t$ with $V_0 = 100$ V to produce a maximum displacement induced current of 1.00 A between the plates?

Exercise:

Problem:

The voltage across a parallel-plate capacitor with area $A = 800\,\text{cm}^2$ and separation $d = 2\,\text{mm}$ varies sinusoidally as $V = (15\,\text{mV})\cos(150t)$, where *t* is in seconds. Find the displacement current between the plates.

Solution:

$I_\text{d} = \left(7.97\ \times\ 10^{-10}\,\text{A}\right)\sin\ (150\,t)$

Exercise:

Problem:

The voltage across a parallel-plate capacitor with area *A* and separation *d* varies with time *t* as $V = at^2$, where *a* is a constant. Find the displacement current between the plates.

Glossary

displacement current	extra term in Maxwell’s equations that is analogous to a real current but accounts for a changing electric field producing a magnetic field, even when the real current is present
Maxwell’s equations	set of four equations that comprise a complete, overarching theory of electromagnetism

Plane Electromagnetic Waves
By the end of this section, you will be able to:

- Describe how Maxwell’s equations predict the relative directions of the electric fields and magnetic fields, and the direction of propagation of plane electromagnetic waves
- Explain how Maxwell’s equations predict that the speed of propagation of electromagnetic waves in free space is exactly the speed of light
- Calculate the relative magnitude of the electric and magnetic fields in an electromagnetic plane wave
- Describe how electromagnetic waves are produced and detected

Mechanical waves travel through a medium such as a string, water, or air. Perhaps the most significant prediction of Maxwell’s equations is the existence of combined electric and magnetic (or electromagnetic) fields that propagate through space as electromagnetic waves. Because Maxwell’s equations hold in free space, the predicted electromagnetic waves, unlike mechanical waves, do not require a medium for their propagation.

A general treatment of the physics of electromagnetic waves is beyond the scope of this textbook. We can, however, investigate the special case of an electromagnetic wave that propagates through free space along the *x*-axis of a given coordinate system.

Electromagnetic Waves in One Direction

An electromagnetic wave consists of an electric field, defined as usual in terms of the force per charge on a stationary charge, and a magnetic field, defined in terms of the force per charge on a moving charge. The electromagnetic field is assumed to be a function of only the *x*-coordinate and time. The *y*-component of the electric field is then written as $E_y\left(x,t\right)$, the *z*-component of the magnetic field as $B_z\left(x,t\right)$, etc. Because we are assuming free space, there are no free charges or currents, so we can set $Q_{\text{in}}=0$ and $I=0$ in Maxwell’s equations.

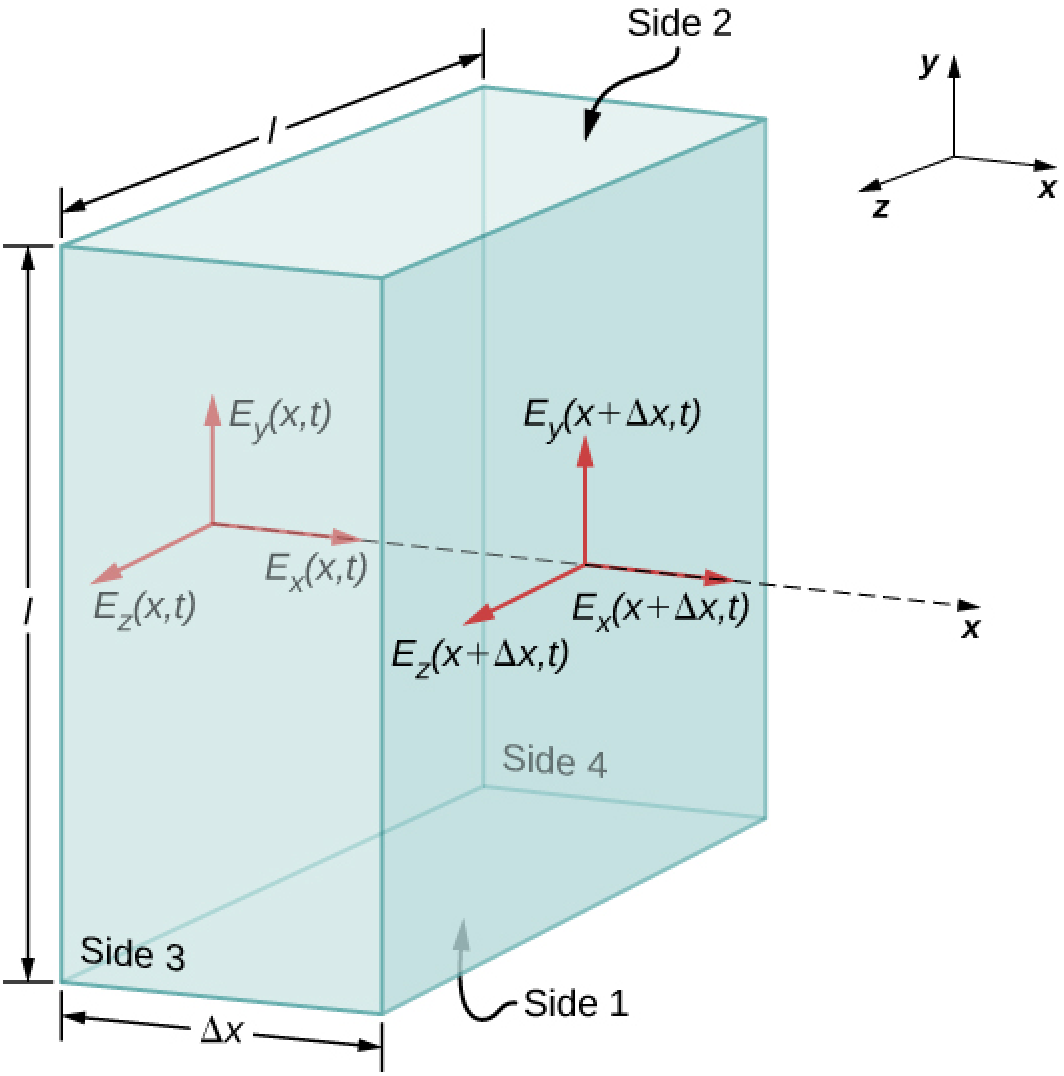
The transverse nature of electromagnetic waves

We examine first what Gauss’s law for electric fields implies about the relative directions of the electric field and the propagation direction in an electromagnetic wave. Assume the Gaussian surface to be the surface of a rectangular box whose cross-section is a square of side *l* and whose third side has length Δx, as shown in [link]. Because the electric field is a function only of *x* and *t*, the *y*-component of the electric field is the same on both the top (labeled Side 2) and bottom (labeled Side 1) of the box, so that these two contributions to the flux cancel. The corresponding argument also holds for the net flux from the *z*-component of the electric field through Sides 3 and 4. Any net flux through the surface therefore comes entirely from the *x*-component of the electric field. Because the electric field has no *y*- or *z*-dependence, $E_x\left(x,t\right)$ is constant over the face of the box with area *A* and has a possibly different value $E_x\left(x+\Delta x,t\right)$ that is constant over the opposite face of the box. Applying Gauss’s law gives

Equation:

$$\text{Net flux} = -E_x\left(x,t\right)A + E_x\left(x+\Delta x,t\right)A = \frac{Q_{\text{in}}}{\varepsilon_0}$$

where $A = l\ \times\ l$ is the area of the front and back faces of the rectangular surface. But the charge enclosed is $Q_{\text{in}} = 0$, so this component’s net flux is also zero, and [link] implies $E_x\left(x,t\right) = E_x\left(x+\Delta x,t\right)$ for any Δx. Therefore, if there is an *x*-component of the electric field, it cannot vary with *x*. A uniform field of that kind would merely be superposed artificially on the traveling wave, for example, by having a pair of parallel-charged plates. Such a component $E_x(x,t)$ would not be part of an electromagnetic wave propagating along the *x*-axis; so $E_x(x,t) = 0$ for this wave. Therefore, the only nonzero components of the electric field are $E_y(x,t)$ and $E_z(x,t)$, perpendicular to the direction of propagation of the wave.


Side 2
y
z
x
l
$E_y(x,t)$
$E_y(x+\Delta x,t)$
$E_x(x,t)$
$E_z(x,t)$
$E_x(x+\Delta x,t)$
$E_z(x+\Delta x,t)$
x
l
Side 4
Side 3
Side 1
Δx

The surface of a rectangular box of dimensions $l \times l \times \Delta x$ is our Gaussian surface. The electric field shown is from an electromagnetic wave propagating along the *x*-axis.

A similar argument holds by substituting *E* for *B* and using Gauss’s law for magnetism instead of Gauss’s law for electric fields. This shows that the *B* field is also perpendicular to the direction of propagation of the wave. The electromagnetic wave is therefore a transverse wave, with its oscillating electric and magnetic fields perpendicular to its direction of propagation.

The speed of propagation of electromagnetic waves

We can next apply Maxwell’s equations to the description given in connection with [link] in the previous section to obtain an equation for the *E* field from the changing *B* field, and for the *B* field from a changing *E* field. We then combine the two equations to show how the changing *E* and *B* fields propagate through space at a speed precisely equal to the speed of light.

First, we apply Faraday’s law over Side 3 of the Gaussian surface, using the path shown in [link]. Because $E_x(x,t)=0$, we have

Equation:

$$\oint \vec{\mathbf{E}}\cdot d\vec{\mathbf{s}}=-E_y\left(x,t\right)l+E_y\left(x+\Delta x,t\right)l.$$

Assuming Δx is small and approximating $E_y\left(x+\Delta x,t\right)$ by

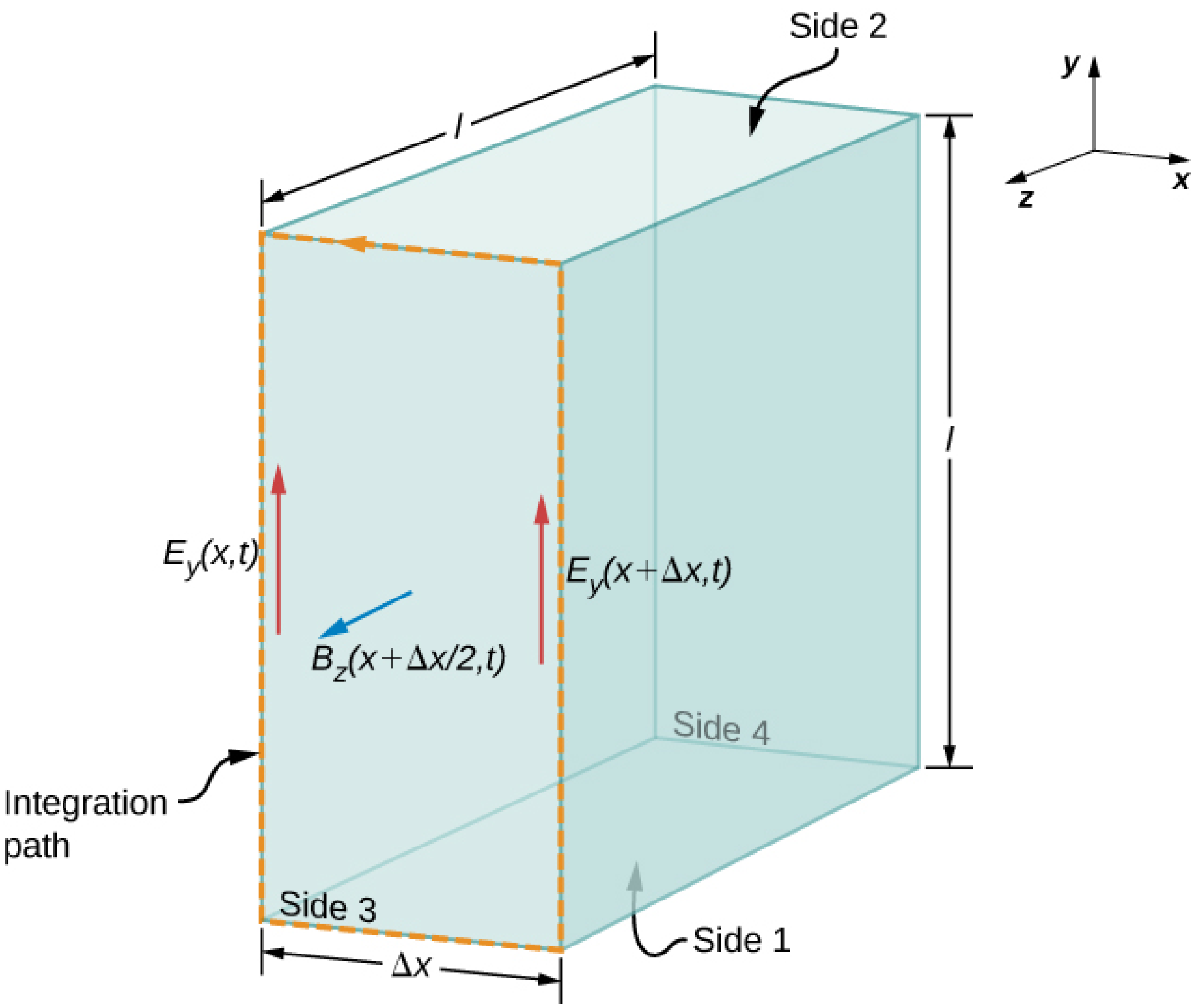
Equation:

$$E_y\left(x+\Delta x,t\right)=E_y\left(x,t\right)+\frac{\partial E_y\left(x,t\right)}{\partial x}\Delta x,$$

we obtain

Equation:

$$\oint \vec{\mathbf{E}}\cdot d\vec{\mathbf{s}}=\frac{\partial E_y\left(x,t\right)}{\partial x}(l\Delta x).$$


Side 2
y
x
z
l
l
$E_y(x,t)$
$E_y(x+\Delta x,t)$
$B_z(x+\Delta x/2,t)$
Side 4
Integration path
Side 3
Side 1
Δx

We apply Faraday's law to the front of the rectangle by evaluating $\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}}$ along the rectangular edge of Side 3 in the direction indicated, taking the B field crossing the face to be approximately its value in the middle of the area traversed.

Because Δx is small, the magnetic flux through the face can be approximated by its value in the center of the area traversed, namely $B_z\left(x+\frac{\Delta x}{2},t\right)$. The flux of the B field through Face 3 is then the B field times the area,

Equation:

$$\oint_S \vec{\mathbf{B}} \cdot \vec{\mathbf{n}} dA = B_z \left(x + \frac{\Delta x}{2}, t \right) (l \Delta x).$$

From Faraday’s law,
Equation:

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d}{dt} \int_S \vec{\mathbf{B}} \cdot \vec{\mathbf{n}} dA.$$

Therefore, from [link] and [link],
Equation:

$$\frac{\partial E_y (x, t)}{\partial x} (l \Delta x) = -\frac{\partial}{\partial t} \left[B_z \left(x + \frac{\Delta x}{2}, t \right) \right] (l \Delta x).$$

Canceling $l \Delta x$ and taking the limit as $\Delta x = 0$, we are left with
Equation:

$$\frac{\partial E_y (x, t)}{\partial x} = -\frac{\partial B_z (x, t)}{\partial t}.$$

We could have applied Faraday’s law instead to the top surface (numbered 2) in [link], to obtain the resulting equation
Equation:

$$\frac{\partial E_z (x, t)}{\partial x} = -\frac{\partial B_y (x, t)}{\partial t}.$$

This is the equation describing the spatially dependent E field produced by the time-dependent B field.

Next we apply the Ampère-Maxwell law (with $I = 0$) over the same two faces (Surface 3 and then Surface 2) of the rectangular box of [link].

Applying [link],

Equation:

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 \varepsilon_0 \left(d/dt \right) \int_S \vec{\mathbf{E}} \cdot \mathbf{n}\, da$$

to Surface 3, and then to Surface 2, yields the two equations

Equation:

$$\frac{\partial B_y\left(x,t\right)}{\partial x} = -\varepsilon_0 \mu_0 \frac{\partial E_z\left(x,t\right)}{\partial t}, \text{ and}$$

Equation:

$$\frac{\partial B_z\left(x,t\right)}{\partial x} = -\varepsilon_0 \mu_0 \frac{\partial E_y\left(x,t\right)}{\partial t}.$$

These equations describe the spatially dependent *B* field produced by the time-dependent *E* field.

We next combine the equations showing the changing *B* field producing an *E* field with the equation showing the changing *E* field producing a *B* field. Taking the derivative of [link] with respect to *x* and using [link] gives

Equation:

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial E_y}{\partial x}\right) = -\frac{\partial}{\partial x}\left(\frac{\partial B_z}{\partial t}\right) = -\frac{\partial}{\partial t}\left(\frac{\partial B_z}{\partial x}\right) = \frac{\partial}{\partial t}\left(\varepsilon_0 \mu_0 \frac{\partial E_y}{\partial t}\right)$$

or

Note:

Equation:

$$\frac{\partial^2 E_y}{\partial x^2} = \varepsilon_0 \mu_0 \frac{\partial^2 E_y}{\partial t^2}.$$

This is the form taken by the general wave equation for our plane wave. Because the equations describe a wave traveling at some as-yet-unspecified speed *c*, we can assume the field components are each functions of *x* – *ct* for the wave traveling in the +*x*-direction, that is,

Equation:

$$E_y(x, t) = f(\xi) \quad \text{where } \xi = x - ct.$$

It is left as a mathematical exercise to show, using the chain rule for differentiation, that [link] and [link] imply

Equation:

$$1 = \varepsilon_0 \mu_0 c^2.$$

The speed of the electromagnetic wave in free space is therefore given in terms of the permeability and the permittivity of free space by

Note:

Equation:

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}.$$

We could just as easily have assumed an electromagnetic wave with field components $E_z(x, t)$ and $B_y(x, t)$. The same type of analysis with [link] and [link] would also show that the speed of an electromagnetic wave is $c = 1/\sqrt{\varepsilon_0 \mu_0}$.

The physics of traveling electromagnetic fields was worked out by Maxwell in 1873. He showed in a more general way than our derivation that electromagnetic waves always travel in free space with a speed given by [link]. If we evaluate the speed $c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$, we find that

Equation:

$$c = \frac{1}{\sqrt{\left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N·m}^2}\right)\left(4\pi \times 10^{-7} \frac{\text{T·m}}{\text{A}}\right)}} = 3.00 \times 10^8 \text{ m/s},$$

which is the speed of light. Imagine the excitement that Maxwell must have felt when he discovered this equation! He had found a fundamental connection between two seemingly unrelated phenomena: electromagnetic fields and light.

Note:
Exercise:

Problem:

Check Your Understanding The wave equation was obtained by (1) finding the E field produced by the changing B field, (2) finding the B field produced by the changing E field, and combining the two results. Which of Maxwell’s equations was the basis of step (1) and which of step (2)?

Solution:

(1) Faraday’s law, (2) the Ampère-Maxwell law

How the *E* and *B* Fields Are Related

So far, we have seen that the rates of change of different components of the *E* and *B* fields are related, that the electromagnetic wave is transverse, and that the wave propagates at speed *c*. We next show what Maxwell's equations imply about the ratio of the *E* and *B* field magnitudes and the relative directions of the *E* and *B* fields.

We now consider solutions to [link] in the form of plane waves for the electric field:

Equation:

$$E_y\left(x,t\right)=E_0\cos\left(kx-\omega t\right).$$

We have arbitrarily taken the wave to be traveling in the +*x*-direction and chosen its phase so that the maximum field strength occurs at the origin at time $t=0$. We are justified in considering only sines and cosines in this way, and generalizing the results, because Fourier's theorem implies we can express any wave, including even square step functions, as a superposition of sines and cosines.

At any one specific point in space, the *E* field oscillates sinusoidally at angular frequency ω between $+E_0$ and $-E_0$, and similarly, the *B* field oscillates between $+B_0$ and $-B_0$. The amplitude of the wave is the maximum value of $E_y(x,t)$. The period of oscillation *T* is the time required for a complete oscillation. The frequency *f* is the number of complete oscillations per unit of time, and is related to the angular frequency ω by $\omega=2\pi f$. The wavelength λ is the distance covered by one complete cycle of the wave, and the wavenumber *k* is the number of wavelengths that fit into a distance of 2π in the units being used. These quantities are related in the same way as for a mechanical wave:

Equation:

$$\omega=2\pi f,\quad f=\frac{1}{T},\quad k=\frac{2\pi}{\lambda},\quad\text{and}\quad c=f\lambda=\omega/k.$$

Given that the solution of E_y has the form shown in [link], we need to determine the B field that accompanies it. From [link], the magnetic field component B_z must obey

Equation:

$$\frac{\partial B_Z}{\partial t} = -\frac{\partial E_y}{\partial x}$$
$$\frac{\partial B_Z}{\partial t} = -\frac{\partial}{\partial x} E_0 \cos(kx - \omega t) = kE_0 \sin(kx - \omega t).$$

Because the solution for the B-field pattern of the wave propagates in the $+x$-direction at the same speed c as the E-field pattern, it must be a function of $k\,(x - ct) = kx - \omega t$. Thus, we conclude from [link] that B_z is

Equation:

$$B_z\,(x,t) = \frac{k}{\omega} E_0 \cos\,(kx - \omega t) = \frac{1}{c} E_0 \cos\,(kx - \omega t).$$

These results may be written as

Equation:

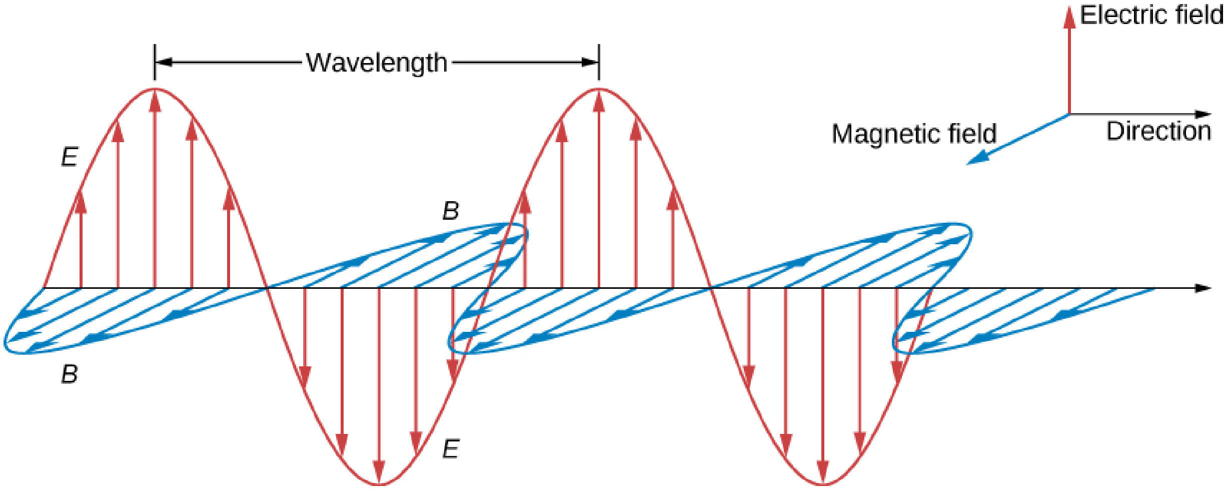
$$E_y\,(x,t) = E_0 \cos\,(kx - \omega t)$$
$$B_z\,(x,t) = B_0 \cos\,(kx - \omega t)$$

Note:

Equation:

$$\frac{E_y}{B_z} = \frac{E_0}{B_0} = c.$$

Therefore, the peaks of the *E* and *B* fields coincide, as do the troughs of the wave, and at each point, the *E* and *B* fields are in the same ratio equal to the speed of light *c*. The plane wave has the form shown in [link].



The plane wave solution of Maxwell's equations has the *B* field directly proportional to the *E* field at each point, with the relative directions shown.

Example:

Calculating *B*-Field Strength in an Electromagnetic Wave

What is the maximum strength of the *B* field in an electromagnetic wave that has a maximum *E*-field strength of 1000 V/m?

Strategy

To find the *B*-field strength, we rearrange [link] to solve for *B*, yielding

Equation:

$$B = \frac{E}{c}.$$

Solution

We are given *E*, and *c* is the speed of light. Entering these into the expression for *B* yields

Equation:

$$B = \frac{1000\ \text{V/m}}{3.00 \times 10^{8}\ \text{m/s}} = 3.33 \times 10^{-6}\ \text{T}.$$

Significance

The *B*-field strength is less than a tenth of Earth’s admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field.

Changing electric fields create relatively weak magnetic fields. The combined electric and magnetic fields can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

Note:

Exercise:

Problem:

Check Your Understanding What conclusions did our analysis of Maxwell’s equations lead to about these properties of a plane electromagnetic wave:

(a) the relative directions of wave propagation, of the *E* field, and of *B* field,

(b) the speed of travel of the wave and how the speed depends on frequency, and

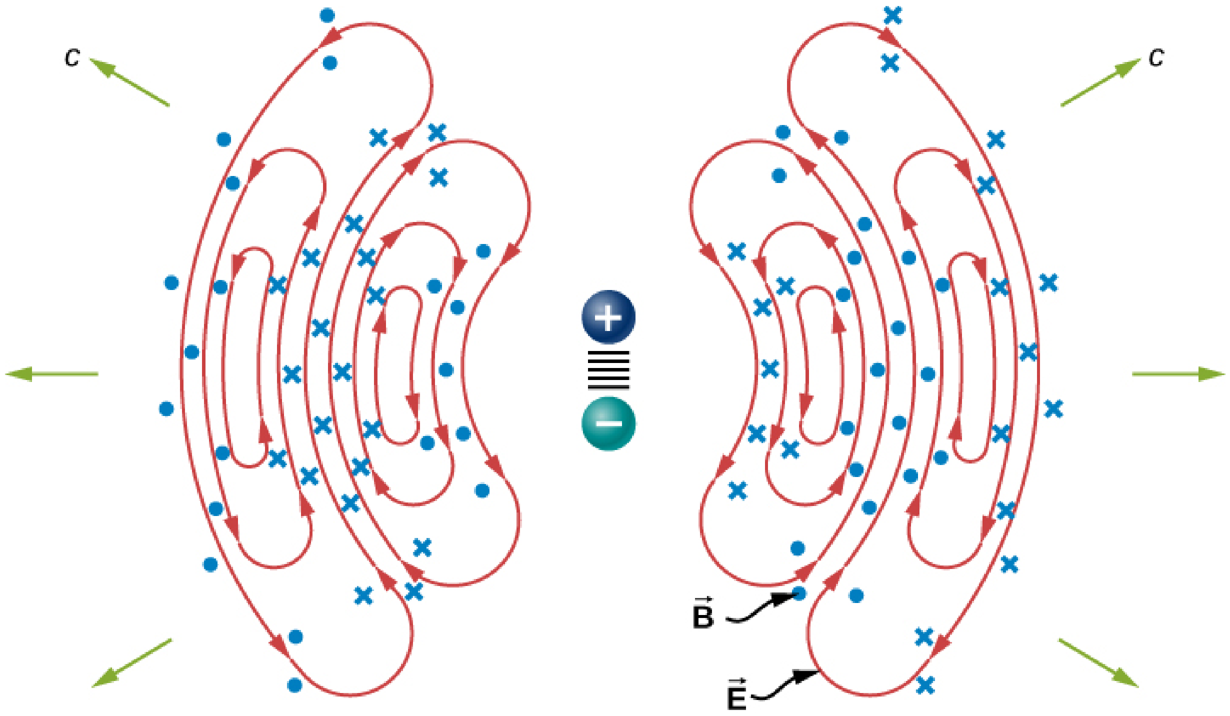
(c) the relative magnitudes of the *E* and *B* fields.

Solution:

a. The directions of wave propagation, of the *E* field, and of *B* field are all mutually perpendicular. b. The speed of the electromagnetic wave is the speed of light $c = 1/\sqrt{\varepsilon_0 \mu_0}$ independent of frequency. c. The ratio of electric and magnetic field amplitudes is $E/B = c$.

Production and Detection of Electromagnetic Waves

A steady electric current produces a magnetic field that is constant in time and which does not propagate as a wave. Accelerating charges, however, produce electromagnetic waves. An electric charge oscillating up and down, or an alternating current or flow of charge in a conductor, emit radiation at the frequencies of their oscillations. The electromagnetic field of a *dipole antenna* is shown in [link]. The positive and negative charges on the two conductors are made to reverse at the desired frequency by the output of a transmitter as the power source. The continually changing current accelerates charge in the antenna, and this results in an oscillating electric field a distance away from the antenna. The changing electric fields produce changing magnetic fields that in turn produce changing electric fields, which thereby propagate as electromagnetic waves. The frequency of this radiation is the same as the frequency of the ac source that is accelerating the electrons in the antenna. The two conducting elements of the dipole antenna are commonly straight wires. The total length of the two wires is typically about one-half of the desired wavelength (hence, the alternative name *half-wave antenna*), because this allows standing waves to be set up and enhances the effectiveness of the radiation.



The oscillatory motion of the charges in a dipole antenna produces electromagnetic radiation.

The electric field lines in one plane are shown. The magnetic field is perpendicular to this plane. This radiation field has cylindrical symmetry around the axis of the dipole. Field lines near the dipole are not shown. The pattern is not at all uniform in all directions. The strongest signal is in directions perpendicular to the axis of the antenna, which would be horizontal if the antenna is mounted vertically. There is zero intensity along the axis of the antenna. The fields detected far from the antenna are from the changing electric and magnetic fields inducing each other and traveling as electromagnetic waves. Far from the antenna, the wave fronts, or surfaces of equal phase for the electromagnetic wave, are almost spherical. Even farther from the antenna, the radiation propagates like electromagnetic plane waves.

The electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving electromagnetic signals works in reverse.

Incoming electromagnetic waves induce oscillating currents in the antenna, each at its own frequency. The radio receiver includes a tuner circuit, whose resonant frequency can be adjusted. The tuner responds strongly to the desired frequency but not others, allowing the user to tune to the desired broadcast. Electrical components amplify the signal formed by the moving electrons. The signal is then converted into an audio and/or video format.

Note:
Use this simulation to broadcast radio waves. Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.

Summary

- Maxwell’s equations predict that the directions of the electric and magnetic fields of the wave, and the wave’s direction of propagation, are all mutually perpendicular. The electromagnetic wave is a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by $c = E/B$, which implies that the magnetic field B is very weak relative to the electric field E.
- Accelerating charges create electromagnetic waves (for example, an oscillating current in a wire produces electromagnetic waves with the same frequency as the oscillation).

Conceptual Questions

Exercise:

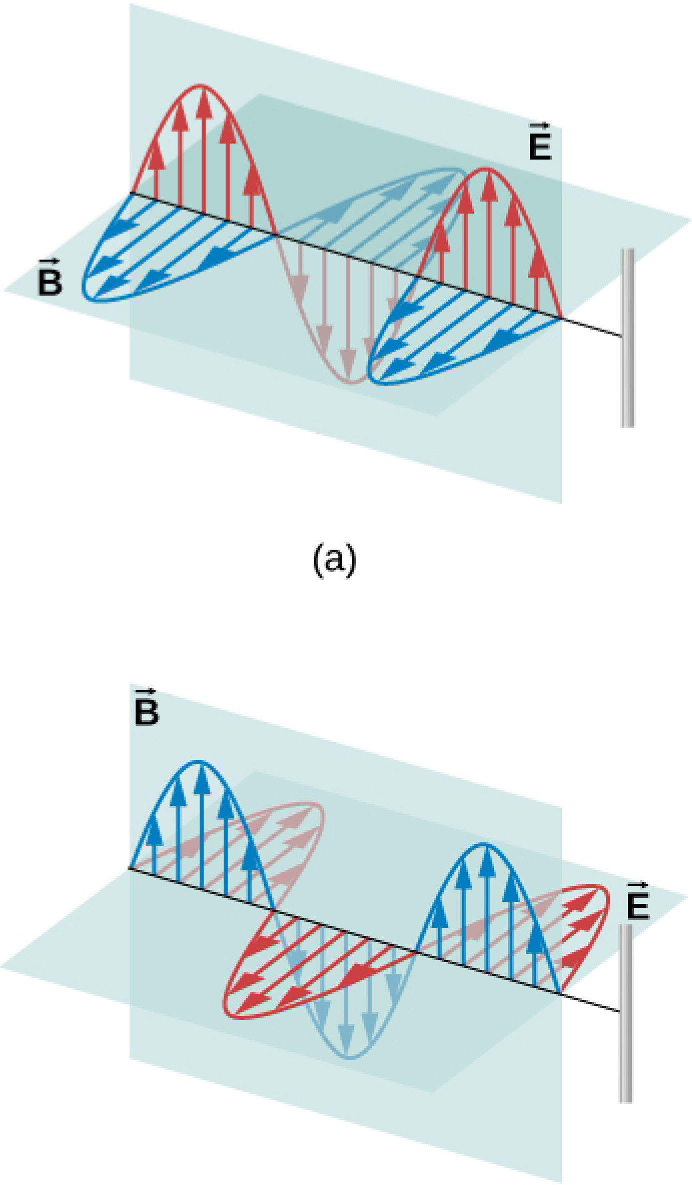
Problem:

If the electric field of an electromagnetic wave is oscillating along the *z*-axis and the magnetic field is oscillating along the *x*-axis, in what possible direction is the wave traveling?

Exercise:

Problem:

In which situation shown below will the electromagnetic wave be more successful in inducing a current in the wire? Explain.


$\vec{E}$
$\vec{B}$
$\vec{B}$
$\vec{E}$

(a)

(b)

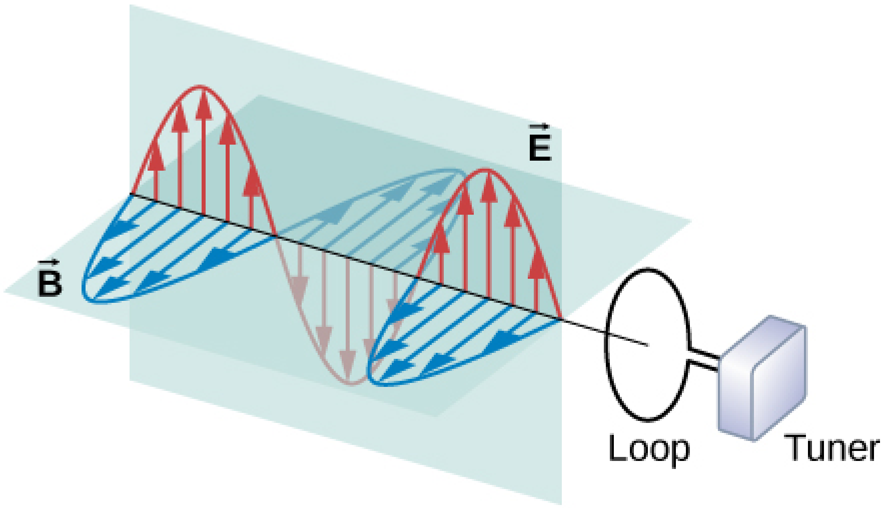
Solution:

in (a), because the electric field is parallel to the wire, accelerating the electrons

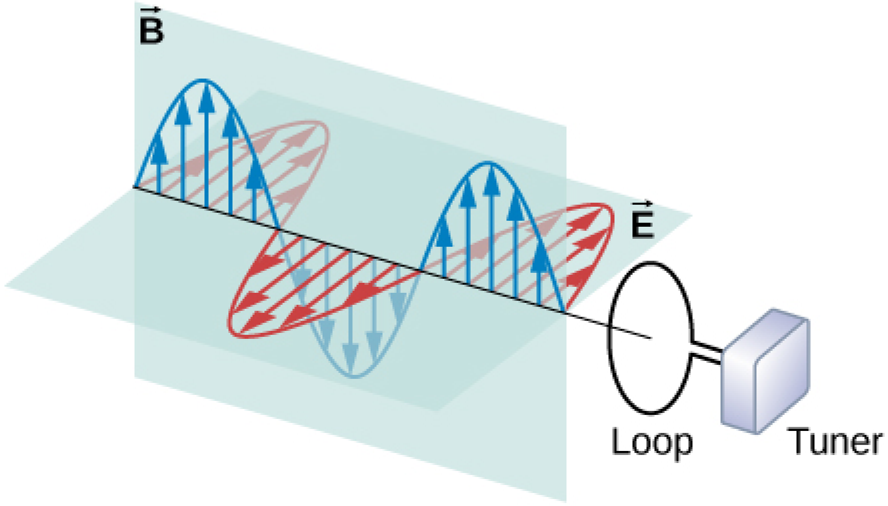
Exercise:

Problem:

In which situation shown below will the electromagnetic wave be more successful in inducing a current in the loop? Explain.


$\vec{E}$
$\vec{B}$
Loop
Tuner

(a)


$\vec{B}$
$\vec{E}$
Loop
Tuner

(b)

Exercise:

Problem:

Under what conditions might wires in a circuit where the current flows in only one direction emit electromagnetic waves?

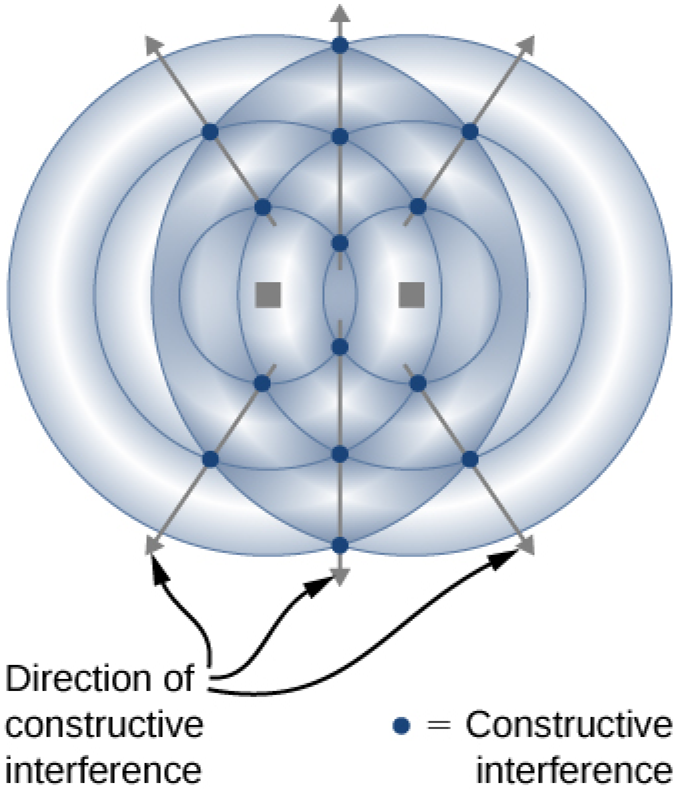
Solution:

A steady current in a dc circuit will not produce electromagnetic waves. If the magnitude of the current varies while remaining in the same direction, the wires will emit electromagnetic waves, for example, if the current is turned on or off.

Exercise:

Problem:

Shown below is the interference pattern of two radio antennas broadcasting the same signal. Explain how this is analogous to the interference pattern for sound produced by two speakers. Could this be used to make a directional antenna system that broadcasts preferentially in certain directions? Explain.


Direction of constructive interference
● = Constructive interference

Problems

Exercise:

Problem:

If the Sun suddenly turned off, we would not know it until its light stopped coming. How long would that be, given that the Sun is 1.496×10^{11} m away?

Solution:

499 s

Exercise:

Problem:

What is the maximum electric field strength in an electromagnetic wave that has a maximum magnetic field strength of $5.00 \times 10^{-4}\,\mathrm{T}$ (about 10 times Earth’s magnetic field)?

Exercise:

Problem:

An electromagnetic wave has a frequency of 12 MHz. What is its wavelength in vacuum?

Solution:

25 m

Exercise:

Problem:

If electric and magnetic field strengths vary sinusoidally in time at frequency 1.00 GHz, being zero at $t = 0$, then $E = E_0 \sin 2\pi ft$ and $B = B_0 \sin 2\pi ft$. (a) When are the field strengths next equal to zero? (b) When do they reach their most negative value? (c) How much time is needed for them to complete one cycle?

Exercise:

Problem:

The electric field of an electromagnetic wave traveling in vacuum is described by the following wave function:
$\vec{\mathbf{E}} = (5.00\ \mathrm{V/m})\cos\left[kx - \left(6.00 \times 10^{9}\ \mathrm{s}^{-1}\right)t + 0.40\right]\hat{\mathbf{j}}$
where *k* is the wavenumber in rad/m, *x* is in m, *t* is in s.
Find the following quantities:
(a) amplitude
(b) frequency
(c) wavelength
(d) the direction of the travel of the wave
(e) the associated magnetic field wave

Solution:

a. 5.00 V/m; b. 9.55×10^{8}Hz; c. 31.4 cm; d. toward the +*x*-axis;
e. $B = \left(1.67 \times 10^{-8}\ \mathrm{T}\right)\cos\left[kx - \left(6 \times 10^{9}\ \mathrm{s}^{-1}\right)t + 0.40\right]\widehat{\mathbf{k}}$

Exercise:

Problem:

A plane electromagnetic wave of frequency 20 GHz moves in the positive *y*-axis direction such that its electric field is pointed along the *z*-axis. The amplitude of the electric field is 10 V/m. The start of time is chosen so that at $t = 0$, the electric field has a value 10 V/m at the origin. (a) Write the wave function that will describe the electric field wave. (b) Find the wave function that will describe the associated magnetic field wave.

Exercise:

Problem:

The following represents an electromagnetic wave traveling in the direction of the positive *y*-axis:

$E_x = 0; E_y = E_0 \cos\ (kx - \omega t); E_z = 0$

$B_x = 0; B_y = 0; B_z = B_0 \cos\ (kx - \omega t)$.

The wave is passing through a wide tube of circular cross-section of radius *R* whose axis is along the *y*-axis. Find the expression for the displacement current through the tube.

Solution:

$I_{\mathrm{d}} = \pi \varepsilon_0 \omega R^2 E_0 \sin\ (kx - \omega t)$

Energy Carried by Electromagnetic Waves

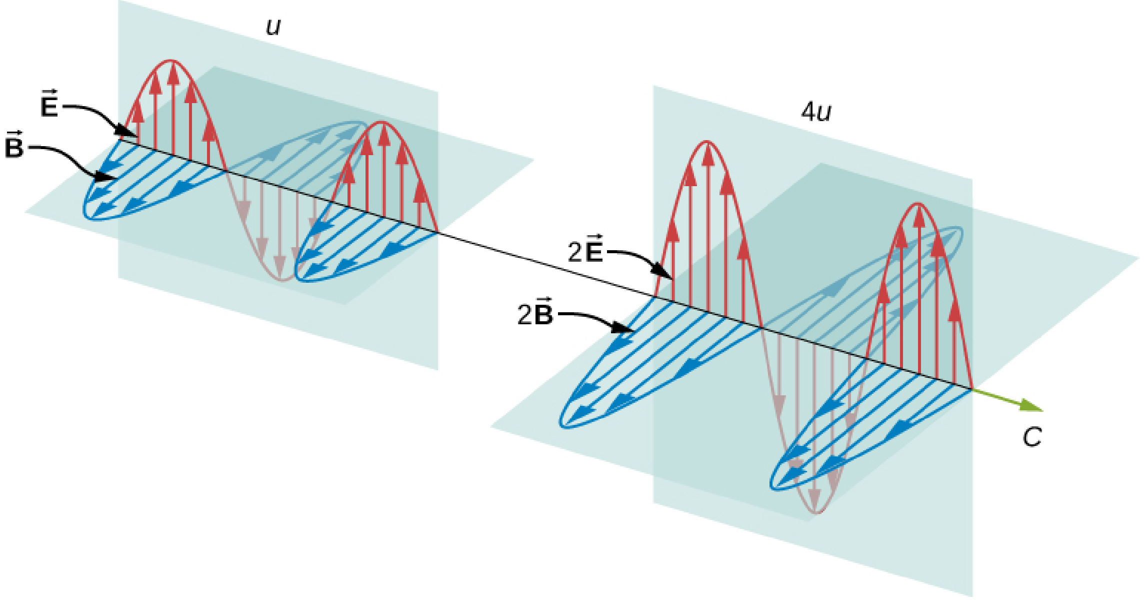
By the end of this section, you will be able to:

- Express the time-averaged energy density of electromagnetic waves in terms of their electric and magnetic field amplitudes
- Calculate the Poynting vector and the energy intensity of electromagnetic waves
- Explain how the energy of an electromagnetic wave depends on its amplitude, whereas the energy of a photon is proportional to its frequency

Anyone who has used a microwave oven knows there is energy in electromagnetic waves. Sometimes this energy is obvious, such as in the warmth of the summer Sun. Other times, it is subtle, such as the unfelt energy of gamma rays, which can destroy living cells.

Electromagnetic waves bring energy into a system by virtue of their electric and magnetic fields. These fields can exert forces and move charges in the system and, thus, do work on them. However, there is energy in an electromagnetic wave itself, whether it is absorbed or not. Once created, the fields carry energy away from a source. If some energy is later absorbed, the field strengths are diminished and anything left travels on.

Clearly, the larger the strength of the electric and magnetic fields, the more work they can do and the greater the energy the electromagnetic wave carries. In electromagnetic waves, the amplitude is the maximum field strength of the electric and magnetic fields ([link]). The wave energy is determined by the wave amplitude.



Energy carried by a wave depends on its amplitude. With electromagnetic waves, doubling the E fields and B fields quadruples the energy density u and the energy flux uc.

For a plane wave traveling in the direction of the positive x-axis with the phase of the wave chosen so that the wave maximum is at the origin at $t = 0$, the electric and magnetic fields obey the equations

Equation:

$$E_y(x,t) = E_0 \cos\,(kx - \omega t)$$
$$B_z(x,t) = B_0 \cos\,(kx - \omega t).$$

The energy in any part of the electromagnetic wave is the sum of the energies of the electric and magnetic fields. This energy per unit volume, or energy density *u*, is the sum of the energy density from the electric field and the energy density from the magnetic field. Expressions for both field energy densities were discussed earlier (u_E in Capacitance and u_B in Inductance). Combining these the contributions, we obtain

Equation:

$$u\,(x,t) = u_E + u_B = \frac{1}{2}\varepsilon_0 E^2 + \frac{1}{2\mu_0}B^2.$$

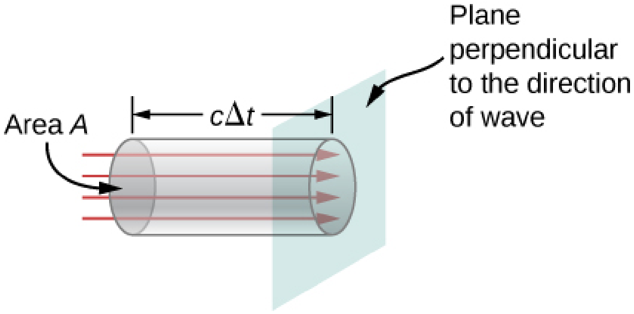
The expression $E = cB = \frac{1}{\sqrt{\varepsilon_0\mu_0}}B$ then shows that the magnetic energy density u_B and electric energy density u_E are equal, despite the fact that changing electric fields generally produce only small magnetic fields. The equality of the electric and magnetic energy densities leads to

Equation:

$$u\,(x,t) = \varepsilon_0 E^2 = \frac{B^2}{\mu_0}.$$

The energy density moves with the electric and magnetic fields in a similar manner to the waves themselves.

We can find the rate of transport of energy by considering a small time interval Δt. As shown in [link], the energy contained in a cylinder of length $c\Delta t$ and cross-sectional area *A* passes through the cross-sectional plane in the interval Δt.



The energy $uAc\Delta t$ contained in the electric and magnetic fields of the electromagnetic wave in the volume $Ac\Delta t$ passes through the area *A* in time Δt.

The energy passing through area *A* in time Δt is

Equation:

$$u\ \times\ \text{volume} = uAc\Delta t.$$

The energy per unit area per unit time passing through a plane perpendicular to the wave, called the energy flux and denoted by S, can be calculated by dividing the energy by the area A and the time interval Δt.
Equation:

$$S = \frac{\text{Energy passing area } A \text{ in time } \Delta t}{A\Delta t} = uc = \varepsilon_0 cE^2 = \frac{1}{\mu_0}EB.$$

More generally, the flux of energy through any surface also depends on the orientation of the surface. To take the direction into account, we introduce a vector $\vec{\mathbf{S}}$, called the **Poynting vector**, with the following definition:

Note:
Equation:

$$\vec{\mathbf{S}} = \frac{1}{\mu_0}\vec{\mathbf{E}}\ \times\ \vec{\mathbf{B}}.$$

The cross-product of $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ points in the direction perpendicular to both vectors. To confirm that the direction of $\vec{\mathbf{S}}$ is that of wave propagation, and not its negative, return to [link]. Note that Lenz's and Faraday's laws imply that when the magnetic field shown is increasing in time, the electric field is greater at x than at $x + \Delta x$. The electric field is decreasing with increasing x at the given time and location. The proportionality between electric and magnetic fields requires the electric field to increase in time along with the magnetic field. This is possible only if the wave is propagating to the right in the diagram, in which case, the relative orientations show that $\vec{\mathbf{S}} = \frac{1}{\mu_0}\vec{\mathbf{E}}\ \times\ \vec{\mathbf{B}}$ is specifically in the direction of propagation of the electromagnetic wave.

The energy flux at any place also varies in time, as can be seen by substituting u from [link] into [link].
Equation:

$$S\left(x,t\right) = c\varepsilon_0 E_0^2 \cos^2\left(kx - \omega t\right)$$

Because the frequency of visible light is very high, of the order of 10^{14} Hz, the energy flux for visible light through any area is an extremely rapidly varying quantity. Most measuring devices, including our eyes, detect only an average over many cycles. The time average of the energy flux is the intensity I of the electromagnetic wave and is the power per unit area. It can be expressed by averaging the cosine function in [link] over one complete cycle, which is the same as time-averaging over many cycles (here, T is one period):
Equation:

$$I=S_{\text{avg}}=c\varepsilon_0E_0^2\frac{1}{T}\int_0^T\cos^2\left(2\pi\frac{t}{T}\right)dt.$$

We can either evaluate the integral, or else note that because the sine and cosine differ merely in phase, the average over a complete cycle for $\cos^2(\xi)$ is the same as for $\sin^2(\xi)$, to obtain
Equation:

$$\langle\cos^2\xi\rangle=\frac{1}{2}\left[\langle\cos^2\xi\rangle+\langle\sin^2\xi\rangle\right]=\frac{1}{2}\langle 1\rangle=\frac{1}{2}.$$

where the angle brackets $\langle\cdots\rangle$ stand for the time-averaging operation. The intensity of light moving at speed c in vacuum is then found to be

Note:
Equation:

$$I=S_{\text{avg}}=\frac{1}{2}c\varepsilon_0E_0^2$$

in terms of the maximum electric field strength E_0, which is also the electric field amplitude. Algebraic manipulation produces the relationship

Note:
Equation:

$$I=\frac{cB_0^2}{2\mu_0}$$

where B_0 is the magnetic field amplitude, which is the same as the maximum magnetic field strength. One more expression for I_{avg} in terms of both electric and magnetic field strengths is useful. Substituting the fact that $cB_0=E_0$, the previous expression becomes

Note:
Equation:

$$I=\frac{E_0B_0}{2\mu_0}.$$

We can use whichever of the three preceding equations is most convenient, because the three equations are really just different versions of the same result: The energy in a wave is related to amplitude squared. Furthermore, because these equations are based on the assumption that the electromagnetic waves are sinusoidal, the peak intensity is twice the average intensity; that is, $I_0 = 2I$.

Example:
A Laser Beam
The beam from a small laboratory laser typically has an intensity of about $1.0\ \times\ 10^{-3}\,\text{W/m}^2$. Assuming that the beam is composed of plane waves, calculate the amplitudes of the electric and magnetic fields in the beam.
Strategy
Use the equation expressing intensity in terms of electric field to calculate the electric field from the intensity.
Solution
From [link], the intensity of the laser beam is
Equation:

$$I = \frac{1}{2}c\varepsilon_0 E_0^2.$$

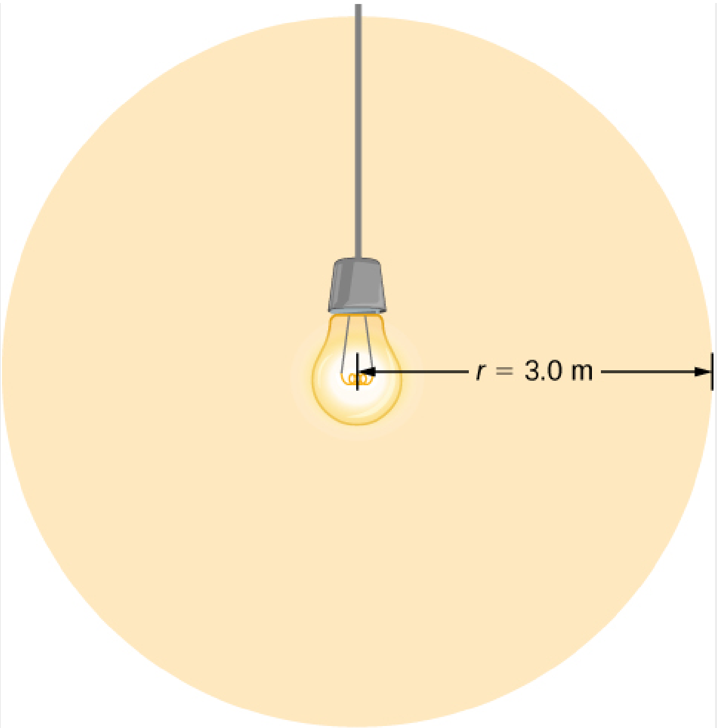
The amplitude of the electric field is therefore
Equation:

$$E_0 = \sqrt{\frac{2}{c\varepsilon_0}I} = \sqrt{\frac{2}{\left(3.00\ \times\ 10^{8}\,\text{m/s}\right)\left(8.85\ \times\ 10^{-12}\,\text{F/m}\right)}\left(1.0\ \times\ 10^{-3}\,\text{W/m}^2\right)} = 0.87\,\text{V/m}.$$

The amplitude of the magnetic field can be obtained from [link]:
Equation:

$$B_0 = \frac{E_0}{c} = 2.9\ \times\ 10^{-9}\,\text{T}.$$

Example:
Light Bulb Fields
A light bulb emits 5.00 W of power as visible light. What are the average electric and magnetic fields from the light at a distance of 3.0 m?
Strategy
Assume the bulb’s power output P is distributed uniformly over a sphere of radius 3.0 m to calculate the intensity, and from it, the electric field.



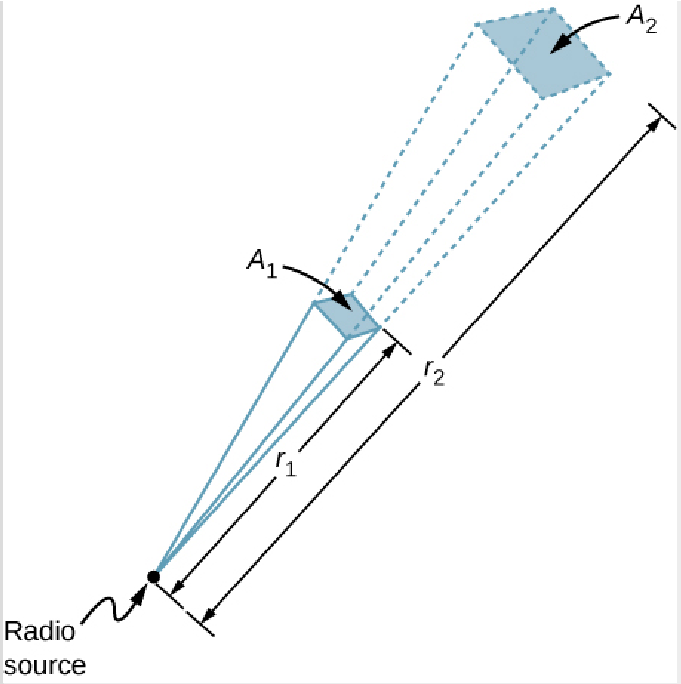
Solution
The power radiated as visible light is then

Equation:

$$\begin{aligned} I &= \frac{P}{4\pi r^2} = \frac{c\varepsilon_0 E_0^2}{2}, \\ E_0 &= \sqrt{2\frac{P}{4\pi r^2 c\varepsilon_0}} = \sqrt{2\frac{5.00\ \text{W}}{4\pi(3.0\ \text{m})^2\left(3.00\times10^8\ \text{m/s}\right)\left(8.85\times10^{-12}\ \text{C}^2/\text{N}\cdot\text{m}^2\right)}} = 5.77\ \text{N/C}, \\ B_0 &= E_0/c = 1.92\ \times\ 10^{-8}\ \text{T}. \end{aligned}$$

Significance
The intensity *I* falls off as the distance squared if the radiation is dispersed uniformly in all directions.

Example:
Radio Range
A 60-kW radio transmitter on Earth sends its signal to a satellite 100 km away ([link]). At what distance in the same direction would the signal have the same maximum field strength if the transmitter's output power were increased to 90 kW?



In three dimensions, a signal spreads over a solid angle as it travels outward from its source.

Strategy

The area over which the power in a particular direction is dispersed increases as distance squared, as illustrated in the figure. Change the power output P by a factor of (90 kW/60 kW) and change the area by the same factor to keep $I = \frac{P}{A} = \frac{c\varepsilon_0 E_0^2}{2}$ the same. Then use the proportion of area A in the diagram to distance squared to find the distance that produces the calculated change in area.

Solution

Using the proportionality of the areas to the squares of the distances, and solving, we obtain from the diagram

Equation:

$$\begin{aligned}\frac{r_2^2}{r_1^2} &= \frac{A_2}{A_1} = \frac{90\text{ W}}{60\text{ W}},\\ r_2 &= \sqrt{\frac{90}{60}}\,(100\,\text{km}) = 122\,\text{km}.\end{aligned}$$

Significance

The range of a radio signal is the maximum distance between the transmitter and receiver that allows for normal operation. In the absence of complications such as reflections from obstacles, the intensity follows an inverse square law, and doubling the range would require multiplying the power by four.

Summary

- The energy carried by any wave is proportional to its amplitude squared. For electromagnetic waves, this means intensity can be expressed as

Equation:

$$I = \frac{c\varepsilon_0 E_0^2}{2}$$

where I is the average intensity in $\mathrm{W/m^2}$ and E_0 is the maximum electric field strength of a continuous sinusoidal wave. This can also be expressed in terms of the maximum magnetic field strength B_0 as

Equation:

$$I = \frac{cB_0^2}{2\mu_0}$$

and in terms of both electric and magnetic fields as

Equation:

$$I = \frac{E_0 B_0}{2\mu_0}.$$

The three expressions for I_{avg} are all equivalent.

Conceptual Questions

Exercise:

Problem:

When you stand outdoors in the sunlight, why can you feel the energy that the sunlight carries, but not the momentum it carries?

Solution:

The amount of energy (about $100\ \mathrm{W/m^2}$) is can quickly produce a considerable change in temperature, but the light pressure (about $3.00 \times 10^{-7}\mathrm{N/m^2}$) is much too small to notice.

Exercise:

Problem:

How does the intensity of an electromagnetic wave depend on its electric field? How does it depend on its magnetic field?

Exercise:

Problem: What is the physical significance of the Poynting vector?

Solution:

It has the magnitude of the energy flux and points in the direction of wave propagation. It gives the direction of energy flow and the amount of energy per area transported per second.

Exercise:

Problem:

A 2.0-mW helium-neon laser transmits a continuous beam of red light of cross-sectional area $0.25\ \text{cm}^2$. If the beam does not diverge appreciably, how would its rms electric field vary with distance from the laser? Explain.

Problems

Exercise:

Problem:

While outdoors on a sunny day, a student holds a large convex lens of radius 4.0 cm above a sheet of paper to produce a bright spot on the paper that is 1.0 cm in radius, rather than a sharp focus. By what factor is the electric field in the bright spot of light related to the electric field of sunlight leaving the side of the lens facing the paper?

Exercise:

Problem:

A plane electromagnetic wave travels northward. At one instant, its electric field has a magnitude of 6.0 V/m and points eastward. What are the magnitude and direction of the magnetic field at this instant?

Solution:

The magnetic field is downward, and it has magnitude $2.00 \times 10^{-8}\text{T}$.

Exercise:

The electric field of an electromagnetic wave is given by $E = \left(6.0 \times 10^{-3}\ \text{V/m}\right) \sin\left[2\pi\left(\frac{x}{18\ \text{m}} - \frac{t}{6.0 \times 10^{-8}\ \text{s}}\right)\right]\hat{\mathbf{j}}$.

Problem: Write the equations for the associated magnetic field and Poynting vector.

Exercise:

Problem:

A radio station broadcasts at a frequency of 760 kHz. At a receiver some distance from the antenna, the maximum magnetic field of the electromagnetic wave detected is $2.15 \times 10^{-11}\text{T}$. (a) What is the maximum electric field? (b) What is the wavelength of the electromagnetic wave?

Solution:

a. $6.45 \times 10^{-3}\ \text{V/m}$; b. 394 m

Exercise:

Problem:

The filament in a clear incandescent light bulb radiates visible light at a power of 5.00 W. Model the glass part of the bulb as a sphere of radius $r_0 = 3.00\,\text{cm}$ and calculate the amount of electromagnetic energy from visible light inside the bulb.

Exercise:

Problem:

At what distance does a 100-W lightbulb produce the same intensity of light as a 75-W lightbulb produces 10 m away? (Assume both have the same efficiency for converting electrical energy in the circuit into emitted electromagnetic energy.)

Solution:

11.5 m

Exercise:

Problem:

An incandescent light bulb emits only 2.6 W of its power as visible light. What is the rms electric field of the emitted light at a distance of 3.0 m from the bulb?

Exercise:

Problem:

A 150-W lightbulb emits 5% of its energy as electromagnetic radiation. What is the magnitude of the average Poynting vector 10 m from the bulb?

Solution:

$5.97\ \times\ 10^{-3}\,\text{W/m}^2$

Exercise:

Problem:

A small helium-neon laser has a power output of 2.5 mW. What is the electromagnetic energy in a 1.0-m length of the beam?

Exercise:

Problem:

At the top of Earth’s atmosphere, the time-averaged Poynting vector associated with sunlight has a magnitude of about $1.4\,\text{kW/m}^2$.
(a) What are the maximum values of the electric and magnetic fields for a wave of this intensity?
(b) What is the total power radiated by the sun? Assume that the Earth is $1.5\ \times\ 10^{11}\text{m}$ from the Sun and that sunlight is composed of electromagnetic plane waves.

Solution:

a. $E_0 = 1027\,\mathrm{V/m}, B_0 = 3.42 \times 10^{-6}\mathrm{T};\ \mathrm{b.}\ 3.96 \times 10^{26}\,\mathrm{W}$

Exercise:

Problem:

The magnetic field of a plane electromagnetic wave moving along the *z* axis is given by $\vec{\mathbf{B}} = B_0\,(\cos kz + \omega t)\hat{\mathbf{j}}$, where $B_0 = 5.00 \times 10^{-10}\,\mathrm{T}$ and $k = 3.14 \times 10^{-2}\,\mathrm{m}^{-1}$. (a) Write an expression for the electric field associated with the wave. (b) What are the frequency and the wavelength of the wave? (c) What is its average Poynting vector?

Exercise:

Problem:

What is the intensity of an electromagnetic wave with a peak electric field strength of 125 V/m?

Solution:

$20.8\,\mathrm{W/m^2}$

Exercise:

Problem:

Assume the helium-neon lasers commonly used in student physics laboratories have power outputs of 0.500 mW. (a) If such a laser beam is projected onto a circular spot 1.00 mm in diameter, what is its intensity? (b) Find the peak magnetic field strength. (c) Find the peak electric field strength.

Exercise:

Problem:

An AM radio transmitter broadcasts 50.0 kW of power uniformly in all directions. (a) Assuming all of the radio waves that strike the ground are completely absorbed, and that there is no absorption by the atmosphere or other objects, what is the intensity 30.0 km away? (*Hint:* Half the power will be spread over the area of a hemisphere.) (b) What is the maximum electric field strength at this distance?

Solution:

a. $4.42 \times 10^{-6}\,\mathrm{W/m^2}$; b. $5.77 \times 10^{-2}\,\mathrm{V/m}$

Exercise:

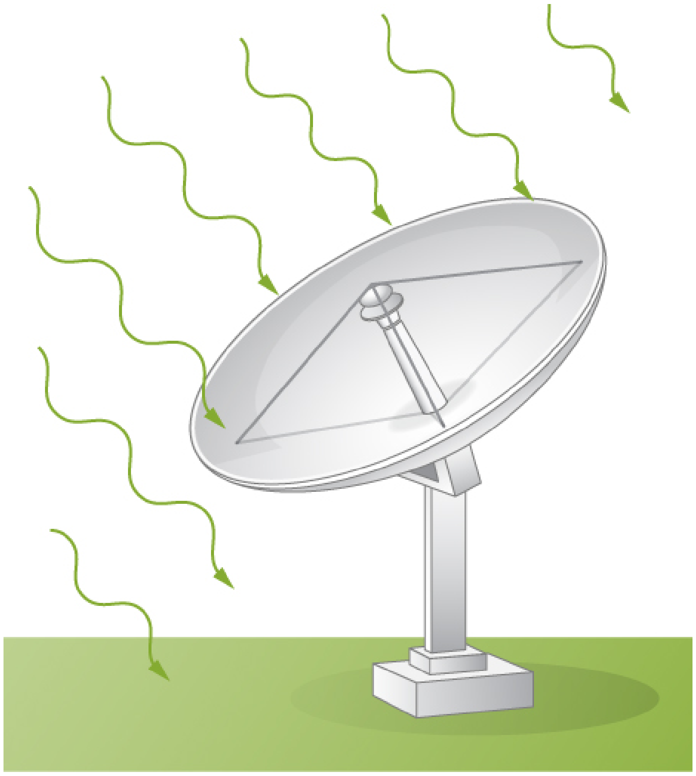
Problem:

Suppose the maximum safe intensity of microwaves for human exposure is taken to be $1.00\,\mathrm{W/m^2}$. (a) If a radar unit leaks 10.0 W of microwaves (other than those sent by its antenna) uniformly in all directions, how far away must you be to be exposed to an intensity considered to be safe? Assume that the power spreads uniformly over the area of a sphere with no complications from absorption or reflection. (b) What is the maximum electric field strength at the safe intensity? (Note that early radar units leaked more than modern ones do. This caused identifiable health problems, such as cataracts, for people who worked near them.)

Exercise:

Problem:

A 2.50-m-diameter university communications satellite dish receives TV signals that have a maximum electric field strength (for one channel) of $7.50\ \mu\text{V/m}$ (see below). (a) What is the intensity of this wave? (b) What is the power received by the antenna? (c) If the orbiting satellite broadcasts uniformly over an area of $1.50 \times 10^{13}\ \text{m}^2$ (a large fraction of North America), how much power does it radiate?



Solution:

a. $7.47 \times 10^{-14}\ \text{W/m}^2$; b. $3.66 \times 10^{-13}\ \text{W}$; c. 1.12 W

Exercise:

Problem:

Lasers can be constructed that produce an extremely high intensity electromagnetic wave for a brief time—called pulsed lasers. They are used to initiate nuclear fusion, for example. Such a laser may produce an electromagnetic wave with a maximum electric field strength of $1.00 \times 10^{11}\ \text{V/m}$ for a time of 1.00 ns. (a) What is the maximum magnetic field strength in the wave? (b) What is the intensity of the beam? (c) What energy does it deliver on an 1.00-mm^2 area?

Glossary

Poynting vector

vector equal to the cross product of the electric-and magnetic fields, that describes the flow of electromagnetic energy through a surface

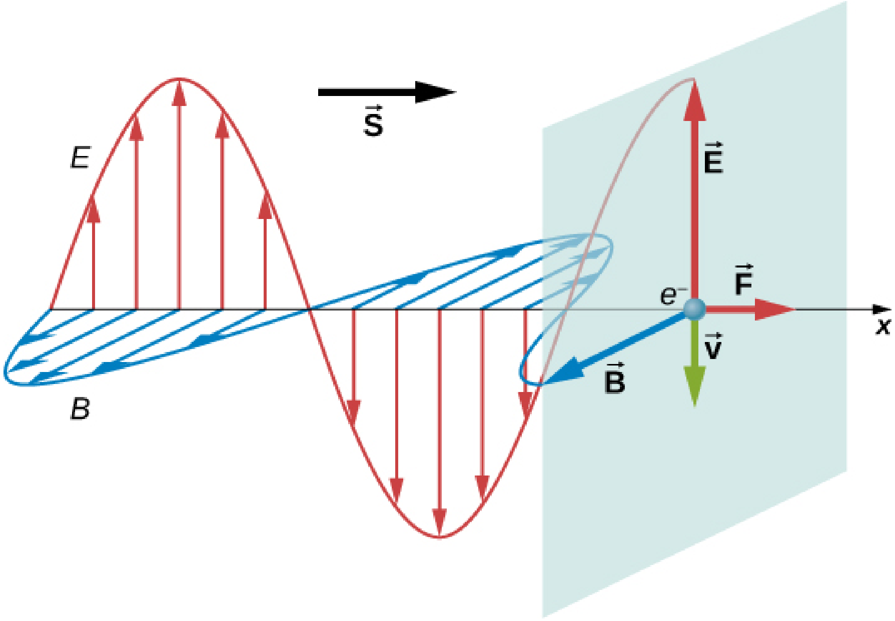
Momentum and Radiation Pressure

By the end of this section, you will be able to:

- Describe the relationship of the radiation pressure and the energy density of an electromagnetic wave
- Explain how the radiation pressure of light, while small, can produce observable astronomical effects

Material objects consist of charged particles. An electromagnetic wave incident on the object exerts forces on the charged particles, in accordance with the Lorentz force, [link]. These forces do work on the particles of the object, increasing its energy, as discussed in the previous section. The energy that sunlight carries is a familiar part of every warm sunny day. A much less familiar feature of electromagnetic radiation is the extremely weak pressure that electromagnetic radiation produces by exerting a force in the direction of the wave. This force occurs because electromagnetic waves contain and transport momentum.

To understand the direction of the force for a very specific case, consider a plane electromagnetic wave incident on a metal in which electron motion, as part of a current, is damped by the resistance of the metal, so that the average electron motion is in phase with the force causing it. This is comparable to an object moving against friction and stopping as soon as the force pushing it stops ([link]). When the electric field is in the direction of the positive *y*-axis, electrons move in the negative *y*-direction, with the magnetic field in the direction of the positive *z*-axis. By applying the right-hand rule, and accounting for the negative charge of the electron, we can see that the force on the electron from the magnetic field is in the direction of the positive *x*-axis, which is the direction of wave propagation. When the *E* field reverses, the *B* field does too, and the force is again in the same direction. Maxwell’s equations together with the Lorentz force equation imply the existence of radiation pressure much more generally than this specific example, however.

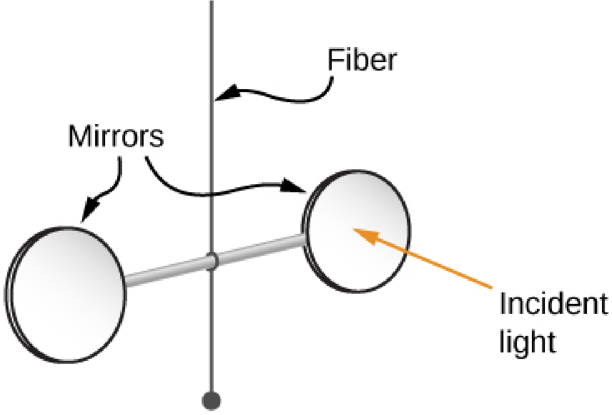


Electric and magnetic fields of an electromagnetic

Electric and magnetic fields of an electromagnetic wave can combine to produce a force in the direction of propagation, as illustrated for the special case of electrons whose motion is highly damped by the resistance of a metal.

Maxwell predicted that an electromagnetic wave carries momentum. An object absorbing an electromagnetic wave would experience a force in the direction of propagation of the wave. The force corresponds to radiation pressure exerted on the object by the wave. The force would be twice as great if the radiation were reflected rather than absorbed.

Maxwell's prediction was confirmed in 1903 by Nichols and Hull by precisely measuring radiation pressures with a torsion balance. The schematic arrangement is shown in [link]. The mirrors suspended from a fiber were housed inside a glass container. Nichols and Hull were able to obtain a small measurable deflection of the mirrors from shining light on one of them. From the measured deflection, they could calculate the unbalanced force on the mirror, and obtained agreement with the predicted value of the force.



Simplified diagram of the central part of the apparatus Nichols and Hull used to precisely measure radiation pressure and confirm Maxwell's prediction.

The **radiation pressure** p_{rad} applied by an electromagnetic wave on a perfectly absorbing surface turns out to be equal to the energy density of the wave:

Equation:

$$p_{\text{rad}} = u \text{ (Perfect absorber).}$$

If the material is perfectly reflecting, such as a metal surface, and if the incidence is along the normal to the surface, then the pressure exerted is twice as much because the momentum direction reverses upon reflection:

Equation:

$$p_{\text{rad}} = 2u \text{ (Perfect reflector)}.$$

We can confirm that the units are right:

Equation:

$$[u] = \frac{\mathrm{J}}{\mathrm{m}^3} = \frac{\mathrm{N} \cdot \mathrm{m}}{\mathrm{m}^3} = \frac{\mathrm{N}}{\mathrm{m}^2} = \text{units of pressure.}$$

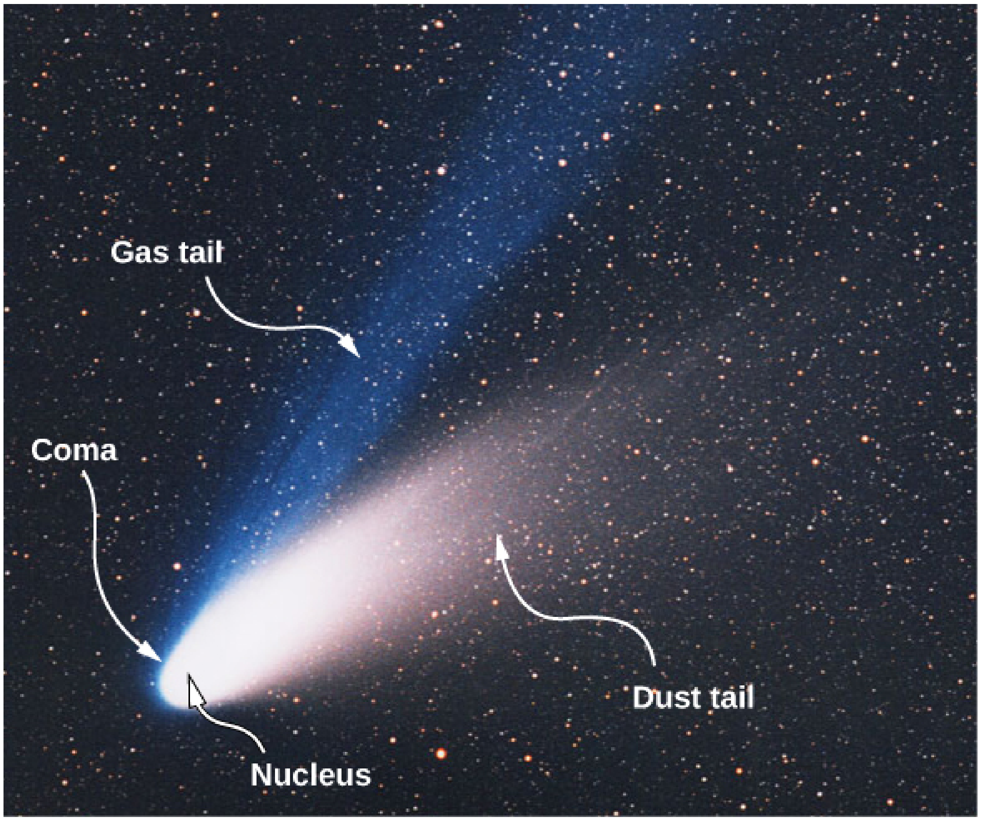
[link] and [link] give the instantaneous pressure, but because the energy density oscillates rapidly, we are usually interested in the time-averaged radiation pressure, which can be written in terms of intensity:

Note:

Equation:

$$p = \langle p_{\text{rad}} \rangle = \begin{cases} I/c & \text{Perfect absorber} \\ 2I/c & \text{Perfect reflector.} \end{cases}$$

Radiation pressure plays a role in explaining many observed astronomical phenomena, including the appearance of comets. Comets are basically chunks of icy material in which frozen gases and particles of rock and dust are embedded. When a comet approaches the Sun, it warms up and its surface begins to evaporate. The *coma* of the comet is the hazy area around it from the gases and dust. Some of the gases and dust form tails when they leave the comet. Notice in [link] that a comet has *two* tails. The *ion tail* (or *gas tail* in [link]) is composed mainly of ionized gases. These ions interact electromagnetically with the solar wind, which is a continuous stream of charged particles emitted by the Sun. The force of the solar wind on the ionized gases is strong enough that the ion tail almost always points directly away from the Sun. The second tail is composed of dust particles. Because the *dust tail* is electrically neutral, it does not interact with the solar wind. However, this tail is affected by the radiation pressure produced by the light from the Sun. Although quite small, this pressure is strong enough to cause the dust tail to be displaced from the path of the comet.



Evaporation of material being warmed by the Sun forms two tails, as shown in this photo of Comet Ison. (credit: modification of work by E. Slawik—ESO)

Example:

Halley’s Comet

On February 9, 1986, Comet Halley was at its closest point to the Sun, about 9.0×10^{10} m from the center of the Sun. The average power output of the Sun is 3.8×10^{26} W.

(a) Calculate the radiation pressure on the comet at this point in its orbit. Assume that the comet reflects all the incident light.

(b) Suppose that a 10-kg chunk of material of cross-sectional area $4.0 \times 10^{-2}\ \text{m}^2$ breaks loose from the comet. Calculate the force on this chunk due to the solar radiation. Compare this force with the gravitational force of the Sun.

Strategy

Calculate the intensity of solar radiation at the given distance from the Sun and use that to calculate the radiation pressure. From the pressure and area, calculate the force.

Solution

a. The intensity of the solar radiation is the average solar power per unit area. Hence, at 9.0×10^{10} m from the center of the Sun, we have

Equation:

$$I = S_{\text{avg}} = \frac{3.8 \times 10^{26}\ \text{W}}{4\pi\left(9.0 \times 10^{10}\ \text{m}\right)^2} = 3.7 \times 10^3\ \text{W/m}^2.$$

Assuming the comet reflects all the incident radiation, we obtain from [link]

Equation:

$$p = \frac{2I}{c} = \frac{2\left(3.7 \times 10^3\ \text{W/m}^2\right)}{3.00 \times 10^8\ \text{m/s}} = 2.5 \times 10^{-5}\ \text{N/m}^2.$$

b. The force on the chunk due to the radiation is

Equation:

$$\begin{aligned} F &= pA = \left(2.5 \times 10^{-5}\ \text{N/m}^2\right)\left(4.0 \times 10^{-2}\ \text{m}^2\right) \\ &= 1.0 \times 10^{-6}\ \text{N}, \end{aligned}$$

whereas the gravitational force of the Sun is

Equation:

$$F_{\text{g}} = \frac{GMm}{r^2} = \frac{\left(6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2\right)\left(2.0 \times 10^{30}\ \text{kg}\right)(10\ \text{kg})}{\left(9.0 \times 10^{10}\ \text{m}\right)^2} = 0.16\ \text{N}.$$

Significance

The gravitational force of the Sun on the chunk is therefore much greater than the force of the radiation.

After Maxwell showed that light carried momentum as well as energy, a novel idea eventually emerged, initially only as science fiction. Perhaps a spacecraft with a large reflecting light sail could use radiation pressure for propulsion. Such a vehicle would not have to carry fuel. It would experience a constant but small force from solar radiation, instead of the short bursts from rocket propulsion. It would accelerate slowly, but by being accelerated continuously, it would eventually reach great speeds. A spacecraft with small total mass and a sail with a large area would be necessary to obtain a usable acceleration.

When the space program began in the 1960s, the idea started to receive serious attention from NASA. The most recent development in light propelled spacecraft has come from a citizen-funded group, the Planetary Society. It is currently testing the use of light sails to propel a small vehicle built from *CubeSats*, tiny satellites that NASA places in orbit for various research projects during space launches intended mainly for other purposes.

The *LightSail* spacecraft shown below ([link]) consists of three *CubeSats* bundled together. It has a total mass of only about 5 kg and is about the size as a loaf of bread. Its sails are made of very thin Mylar and open after launch to have a surface area of $32\,\mathrm{m}^2$.

Two small *CubeSat* satellites deployed from the International Space Station in May, 2016. The solar sails open out when the CubeSats are far enough away from the Station. (credit: modification of work by NASA)

Note:
The first *LightSail* spacecraft was launched in 2015 to test the sail deployment system. It was placed in low-earth orbit in 2015 by hitching a ride on an Atlas 5 rocket launched for an unrelated mission. The test was successful, but the low-earth orbit allowed too much drag on the spacecraft to accelerate it by sunlight. Eventually, it burned in the atmosphere, as expected. The next Planetary Society's *LightSail* solar sailing spacecraft is scheduled for

2016. An illustration of the spacecraft, as it is expected to appear in flight, can be seen on the Planetary Society’s website.

Example:

LightSail Acceleration

The intensity of energy from sunlight at a distance of 1 AU from the Sun is $1370\ \mathrm{W/m^2}$. The *LightSail* spacecraft has sails with total area of $32\ \mathrm{m^2}$ and a total mass of 5.0 kg. Calculate the maximum acceleration LightSail spacecraft could achieve from radiation pressure when it is about 1 AU from the Sun.

Strategy

The maximum acceleration can be expected when the sail is opened directly facing the Sun. Use the light intensity to calculate the radiation pressure and from it, the force on the sails. Then use Newton’s second law to calculate the acceleration.

Solution

The radiation pressure is

Equation:

$$F = pA = 2uA = \frac{2I}{c}A = \frac{2\left(1370\ \mathrm{W/m^2}\right)\left(32\ \mathrm{m^2}\right)}{\left(3.00\ \times\ 10^{8}\ \mathrm{m/s}\right)} = 2.92\ \times\ 10^{-4}\ \mathrm{N}.$$

The resulting acceleration is

Equation:

$$a = \frac{F}{m} = \frac{2.92\ \times\ 10^{-4}\ \mathrm{N}}{5.0\ \mathrm{kg}} = 5.8\ \times\ 10^{-5}\ \mathrm{m/s^2}.$$

Significance

If this small acceleration continued for a year, the craft would attain a speed of 1829 m/s, or 6600 km/h.

Note:

Exercise:

Problem:

Check Your Understanding How would the speed and acceleration of a radiation-propelled spacecraft be affected as it moved farther from the Sun on an interplanetary space flight?

Solution:

Its acceleration would decrease because the radiation force is proportional to the intensity of light from the Sun, which decreases with distance. Its speed, however, would not change except for the effects of gravity from the Sun and planets.

Summary

- Electromagnetic waves carry momentum and exert radiation pressure.
- The radiation pressure of an electromagnetic wave is directly proportional to its energy density.
- The pressure is equal to twice the electromagnetic energy intensity if the wave is reflected and equal to the incident energy intensity if the wave is absorbed.

Conceptual Questions

Exercise:

Problem:

Why is the radiation pressure of an electromagnetic wave on a perfectly reflecting surface twice as large as the pressure on a perfectly absorbing surface?

Solution:

The force on a surface acting over time Δt is the momentum that the force would impart to the object. The momentum change of the light is doubled if the light is reflected back compared with when it is absorbed, so the force acting on the object is twice as great.

Exercise:

Problem:

Why did the early Hubble Telescope photos of Comet Ison approaching Earth show it to have merely a fuzzy coma around it, and not the pronounced double tail that developed later (see below)?

(credit: modification of work by NASA, ESA, J.-Y. Li (Planetary Science Institute), and the Hubble Comet ISON Imaging Science Team)

Exercise:

Problem:

(a) If the electric field and magnetic field in a sinusoidal plane wave were interchanged, in which direction relative to before would the energy propagate?
(b) What if the electric and the magnetic fields were both changed to their negatives?

Solution:

a. According to the right hand rule, the direction of energy propagation would reverse.
b. This would leave the vector $\vec{\mathbf{S}}$, and therefore the propagation direction, the same.

Problems

Exercise:

Problem:

A 150-W lightbulb emits 5% of its energy as electromagnetic radiation. What is the radiation pressure on an absorbing sphere of radius 10 m that surrounds the bulb?

Solution:

$1.99\ \times\ 10^{-11}\,\mathrm{N/m^2}$

Exercise:

Problem:

What pressure does light emitted uniformly in all directions from a 100-W incandescent light bulb exert on a mirror at a distance of 3.0 m, if 2.6 W of the power is emitted as visible light?

Exercise:

Problem:

A microscopic spherical dust particle of radius $2\ \mu\mathrm{m}$ and mass $10\ \mu\mathrm{g}$ is moving in outer space at a constant speed of 30 cm/sec. A wave of light strikes it from the opposite direction of its motion and gets absorbed. Assuming the particle decelerates uniformly to zero speed in one second, what is the average electric field amplitude in the light?

Solution:

$$\begin{aligned} F &= ma = (p)\left(\pi r^2\right),\ p = \tfrac{ma}{\pi r^2} = \tfrac{\varepsilon_0}{2} E_0^2 \\ E_0 &= \sqrt{\frac{2ma}{\varepsilon_0 \pi r^2}} = \sqrt{\frac{2\left(10^{-8}\,\mathrm{kg}\right)\left(0.30\,\mathrm{m/s^2}\right)}{\left(8.854\times10^{-12}\,\mathrm{C^2/N{\cdot}m^2}\right)(\pi)\left(2\times10^{-6}\,\mathrm{m}\right)^2}} \\ E_0 &= 7.34\ \times\ 10^{6}\,\mathrm{V/m} \end{aligned}$$

Exercise:

Problem:

A Styrofoam spherical ball of radius 2 mm and mass $20\ \mu\mathrm{g}$ is to be suspended by the radiation pressure in a vacuum tube in a lab. How much intensity will be required if the light is completely absorbed the ball?

Exercise:

Problem:

Suppose that $\vec{\mathbf{S}}_{\text{avg}}$ for sunlight at a point on the surface of Earth is $900\ \text{W/m}^2$. (a) If sunlight falls perpendicularly on a kite with a reflecting surface of area $0.75\ \text{m}^2$, what is the average force on the kite due to radiation pressure? (b) How is your answer affected if the kite material is black and absorbs all sunlight?

Solution:

a. $4.50\ \times\ 10^{-6}\ \text{N}$; b. it is reduced to half the pressure, $2.25\ \times\ 10^{-6}\ \text{N}$

Exercise:

Problem:

Sunlight reaches the ground with an intensity of about $1.0\ \text{kW/m}^2$. A sunbather has a body surface area of $0.8\ \text{m}^2$ facing the sun while reclining on a beach chair on a clear day. (a) how much energy from direct sunlight reaches the sunbather's skin per second? (b) What pressure does the sunlight exert if it is absorbed?

Exercise:

Problem:

Suppose a spherical particle of mass *m* and radius *R* in space absorbs light of intensity *I* for time *t*. (a) How much work does the radiation pressure do to accelerate the particle from rest in the given time it absorbs the light? (b) How much energy carried by the electromagnetic waves is absorbed by the particle over this time based on the radiant energy incident on the particle?

Solution:

a. $W = \frac{1}{2}\ \frac{\pi^2 r^4}{mc^2} I^2 t^2$; b. $E = \pi r^2 It$

Glossary

radiation pressure
force divided by area applied by an electromagnetic wave on a surface

The Electromagnetic Spectrum

By the end of this section, you will be able to:

- Explain how electromagnetic waves are divided into different ranges, depending on wavelength and corresponding frequency
- Describe how electromagnetic waves in different categories are produced
- Describe some of the many practical everyday applications of electromagnetic waves

Electromagnetic waves have a vast range of practical everyday applications that includes such diverse uses as communication by cell phone and radio broadcasting, WiFi, cooking, vision, medical imaging, and treating cancer. In this module, we discuss how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on. We also summarize some of the main applications for each range.

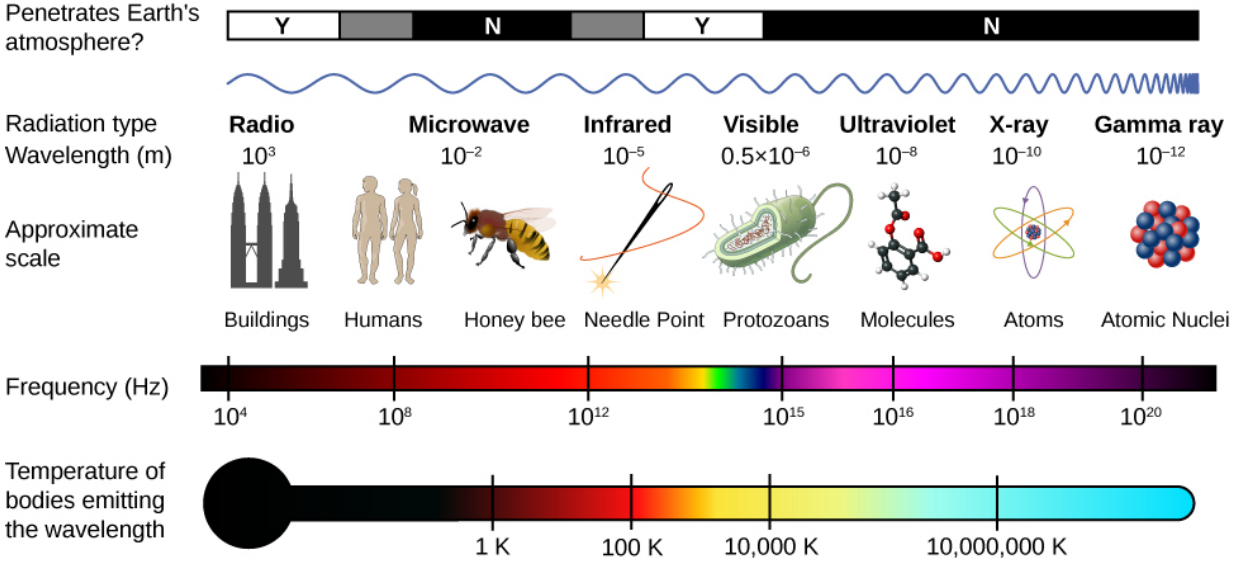
The different categories of electromagnetic waves differ in their wavelength range, or equivalently, in their corresponding frequency ranges. Their properties change smoothly from one frequency range to the next, with different applications in each range. A brief overview of the production and utilization of electromagnetic waves is found in [link].

Type of wave	**Production**	**Applications**	**Issues**
Radio	Accelerating charges	Communications Remote controls MRI	Requires control for band use
Microwaves	Accelerating charges and thermal agitation	Communications Ovens Radar Cell phone use	
Infrared	Thermal agitation and electronic transitions	Thermal imaging Heating	Absorbed by atmosphere Greenhouse effect

Type of wave	Production	Applications	Issues
Visible light	Thermal agitation and electronic transitions	Photosynthesis Human vision	
Ultraviolet	Thermal agitation and electronic transitions	Sterilization Vitamin D production	Ozone depletion Cancer causing
X-rays	Inner electronic transitions and fast collisions	Security Medical diagnosis Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine Security Medical diagnosis Cancer therapy	Cancer causing Radiation damage

Electromagnetic Waves

The relationship $c = f\lambda$ between frequency f and wavelength λ applies to all waves and ensures that greater frequency means smaller wavelength. [link] shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum.


Penetrates Earth's atmosphere?
Y
N
Y
N
Radiation type
Wavelength (m)
Radio
10^{3}
Microwave
10^{-2}
Infrared
10^{-5}
Visible
0.5×10^{-6}
Ultraviolet
10^{-8}
X-ray
10^{-10}
Gamma ray
10^{-12}
Approximate scale
Buildings
Humans
Honey bee
Needle Point
Protozoans
Molecules
Atoms
Atomic Nuclei
Frequency (Hz)
10^{4}
10^{8}
10^{12}
10^{15}
10^{16}
10^{18}
10^{20}
Temperature of bodies emitting the wavelength
1 K
100 K
10,000 K
10,000,000 K

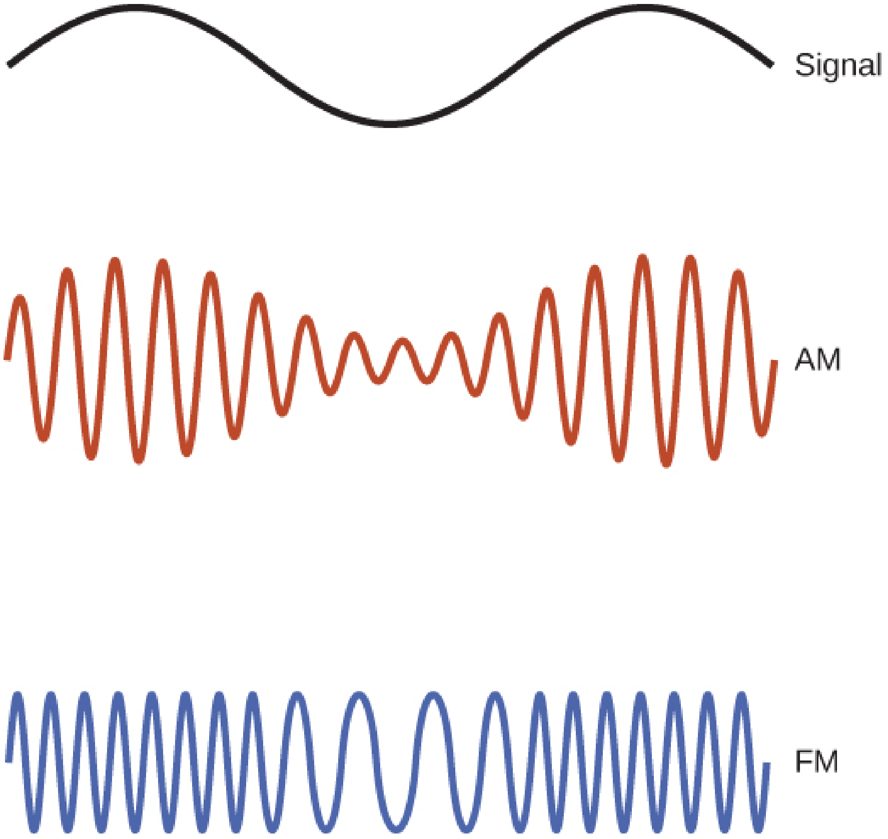
The electromagnetic spectrum, showing the major categories of electromagnetic waves.

Radio Waves

The term **radio waves** refers to electromagnetic radiation with wavelengths greater than about 0.1 m. Radio waves are commonly used for audio communications (i.e., for radios), but the term is used for electromagnetic waves in this range regardless of their application. Radio waves typically result from an alternating current in the wires of a broadcast antenna. They cover a very broad wavelength range and are divided into many subranges, including microwaves, electromagnetic waves used for AM and FM radio, cellular telephones, and TV signals.

There is no lowest frequency of radio waves, but ELF waves, or “extremely low frequency” are among the lowest frequencies commonly encountered, from 3 Hz to 3 kHz. The accelerating charge in the ac currents of electrical power lines produce electromagnetic waves in this range. ELF waves are able to penetrate sea water, which strongly absorbs electromagnetic waves of higher frequency, and therefore are useful for submarine communications.

In order to use an electromagnetic wave to transmit information, the amplitude, frequency, or phase of the wave is *modulated*, or varied in a controlled way that encodes the intended information into the wave. In AM radio transmission, the amplitude of the wave is modulated to mimic the vibrations of the sound being conveyed. Fourier’s theorem implies that the modulated AM wave amounts to a superposition of waves covering some narrow frequency range. Each AM station is assigned a specific carrier frequency that, by international agreement, is allowed to vary by $\pm 5\text{ kHz}$. In FM radio transmission, the frequency of the wave is modulated to carry this information, as illustrated in [link], and the frequency of each station is allowed to use 100 kHz on each side of its carrier frequency. The electromagnetic wave produces a current in a receiving antenna, and the radio or television processes the signal to produce the sound and any image. The higher the frequency of the radio wave used to carry the data, the greater the detailed variation of the wave that can be carried by modulating it over each time unit, and the more data that can be transmitted per unit of time. The assigned frequencies for AM broadcasting are 540 to 1600 kHz, and for FM are 88 MHz to108 MHz.



Electromagnetic waves are used to carry communications signals by varying the wave’s amplitude (AM), its frequency (FM), or its phase.

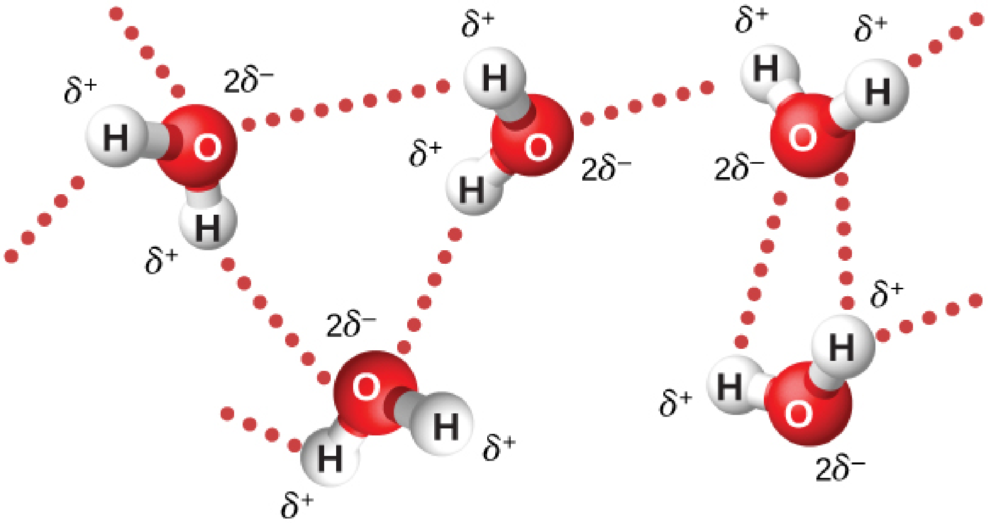
Cell phone conversations, and television voice and video images are commonly transmitted as digital data, by converting the signal into a sequence of binary ones and zeros. This allows clearer data transmission when the signal is weak, and allows using computer algorithms to compress the digital data to transmit more data in each frequency range. Computer data as well is transmitted as a sequence of binary ones and zeros, each one or zero constituting one bit of data.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9Hz to nearly 10^{12}Hz. Their high frequencies correspond to short wavelengths compared with other radio waves—hence the name “microwave.” Microwaves also occur naturally as the cosmic background radiation left over from the origin of the universe. Along with other ranges of electromagnetic waves, they are part of the radiation that any object above absolute zero emits and absorbs because of **thermal agitation**, that is, from the thermal motion of its atoms and molecules.

Most satellite-transmitted information is carried on microwaves. **Radar** is a common application of microwaves. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds, aircraft, or even the surface of Venus.

Microwaves of 2.45 GHz are commonly used in microwave ovens. The electrons in a water molecule tend to remain closer to the oxygen nucleus than the hydrogen nuclei ([link]). This creates two separated centers of equal and opposite charges, giving the molecule a dipole moment (see Electric Field). The oscillating electric field of the microwaves inside the oven exerts a torque that tends to align each molecule first in one direction and then in the other, with the motion of each molecule coupled to others around it. This pumps energy into the continual thermal motion of the water to heat the food. The plate under the food contains no water, and remains relatively unheated.



The oscillating electric field in a microwave oven exerts a torque on water molecules because of their dipole moment, and the torque reverses direction 4.90×10^9 times per second. Interactions between the molecules distributes the energy being pumped into them. The δ^+ and δ^- denote the charge distribution on the molecules.

The microwaves in a microwave oven reflect off the walls of the oven, so that the superposition of waves produces standing waves, similar to the standing waves of a vibrating guitar or violin string (see Normal Modes of a Standing Sound Wave). A rotating fan acts as a stirrer by reflecting the microwaves in different directions, and food turntables, help spread out the hot spots.

Example:

Why Microwave Ovens Heat Unevenly

How far apart are the hotspots in a 2.45-GHz microwave oven?

Strategy

Consider the waves along one direction in the oven, being reflected at the opposite wall from where they are generated.

Solution

The antinodes, where maximum intensity occurs, are half the wavelength apart, with separation

Equation:

$$d = \frac{1}{2}\lambda = \frac{1}{2}\,\frac{c}{f} = \frac{3.00\ \times\ 10^8\,\text{m/s}}{2\left(2.45\ \times\ 10^9\ \text{Hz}\right)} = 6.02\,\text{cm}.$$

Significance

The distance between the hot spots in a microwave oven are determined by the wavelength of the microwaves.

A cell phone has a radio receiver and a weak radio transmitter, both of which can quickly tune to hundreds of specifically assigned microwave frequencies. The low intensity of the transmitted signal gives it an intentionally limited range. A ground-based system links the phone to only to the broadcast tower assigned to the specific small area, or cell, and smoothly transitions its connection to the next cell when the signal reception there is the stronger one. This enables a cell phone to be used while changing location.

Microwaves also provide the WiFi that enables owners of cell phones, laptop computers, and similar devices to connect wirelessly to the Internet at home and at coffee shops and airports. A wireless WiFi router is a device that exchanges data over the Internet through the cable or another connection, and uses microwaves to exchange the data wirelessly with devices such as cell phones and computers. The term WiFi itself refers to the standards followed in modulating and analyzing the microwaves so that wireless routers and devices from different manufacturers work compatibly with one another. The computer data in each direction consist of sequences of binary zeros and ones, each corresponding to a binary bit. The microwaves are in the range of 2.4 GHz to 5.0 GHz range.

Other wireless technologies also use microwaves to provide everyday communications between devices. Bluetooth developed alongside WiFi as a standard for radio communication in the 2.4-GHz range between nearby devices, for example, to link to headphones and audio earpieces to devices such as radios, or a driver’s cell phone to a hands-free device to allow answering phone calls without fumbling directly with the cell phone.

Microwaves find use also in radio tagging, using RFID (radio frequency identification) technology. Examples are RFID tags attached to store merchandize, transponder for toll booths use attached to the windshield of a car, or even a chip embedded into a pet’s skin. The device responds to a microwave signal by emitting a signal of its own with encoded information, allowing stores to quickly ring up items at their cash registers, drivers to charge tolls to their account without stopping, and lost pets to be reunited with their owners. NFC (near field communication) works similarly, except it is much shorter range. Its mechanism of interaction is the induced magnetic field at microwave frequencies between two coils. Cell phones that have NFC capability and the right software can supply information for purchases using the cell phone instead of a physical credit card. The very short range of the data transfer is a desired security feature in this case.

Infrared Radiation

The boundary between the microwave and infrared regions of the electromagnetic spectrum is not well defined (see [link]). **Infrared radiation** is generally produced by thermal motion, and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation. About half of the solar energy arriving at Earth is in the infrared region, with most of the rest in the visible part of the spectrum. About 23% of the solar energy is absorbed in the atmosphere, about 48% is absorbed at Earth’s surface, and about 29% is reflected back into space.[footnote] http://earthobservatory.nasa.gov/Features/EnergyBalance/page4.php

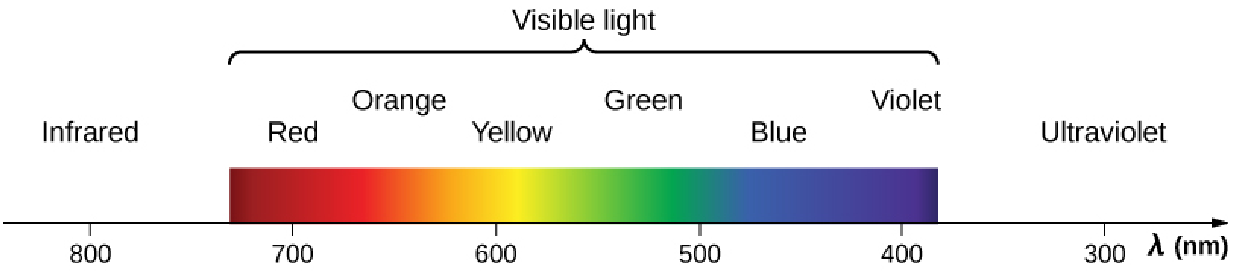
The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means “below red.” Water molecules rotate and vibrate particularly well at infrared frequencies. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, including those called *quartz heaters*, to preferentially warm us because we absorb infrared better than our surroundings.

The familiar handheld “remotes” for changing channels and settings on television sets often transmit their signal by modulating an infrared beam. If you try to use a TV remote without the infrared emitter being in direct line of sight with the infrared detector, you may find the television not responding. Some remotes use Bluetooth instead and reduce this annoyance.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum between about 400 nm and about 750 nm to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions.

Red light has the lowest frequencies and longest wavelengths, whereas violet has the highest frequencies and shortest wavelengths ([link]). Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the sun yellowish in appearance.



A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum in which they are embedded. We enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis uses parts of the visible spectrum to make sugars.

Ultraviolet Radiation

Ultraviolet means “above violet.” The electromagnetic frequencies of **ultraviolet radiation (UV)** extend upward from violet, the highest-frequency visible light. The highest-frequency ultraviolet overlaps with the lowest-frequency X-rays. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies. Ultraviolet is produced by atomic and molecular motions and electronic transitions.

UV radiation from the Sun is broadly subdivided into three wavelength ranges: UV-A (320–400 nm) is the lowest frequency, then UV-B (290–320 nm) and UV-C (220–290 nm). Most UV-B and all UV-C are absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching Earth's surface is UV-A.

Sunburn is caused by large exposures to UV-B and UV-C, and repeated exposure can increase the likelihood of skin cancer. The tanning response is a defense mechanism in which the body produces pigments in inert skin layers to reduce exposure of the living cells below.

As examined in a later chapter, the shorter the wavelength of light, the greater the energy change of an atom or molecule that absorbs the light in an electronic transition. This makes short-wavelength ultraviolet light damaging to living cells. It also explains why ultraviolet radiation is better able than visible light to cause some materials to glow, or *fluoresce*.

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin results from exposure to UV-B radiation, generally from sunlight. Several studies suggest vitamin D deficiency is associated with the development of a range of cancers (prostate, breast, colon), as well as osteoporosis. Low-intensity ultraviolet has applications such as providing the energy to cause certain dyes to fluoresce and emit visible light, for example, in printed money to display hidden watermarks as counterfeit protection.

X-Rays

X-rays have wavelengths from about $10^{-8}\text{m to }10^{-12}\text{m}$. They have shorter wavelengths, and higher frequencies, than ultraviolet, so that the energy they transfer at an atomic level is greater. As a result, X-rays have adverse effects on living cells similar to those of ultraviolet radiation, but they are more penetrating. Cancer and genetic defects can be induced by X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained.

Gamma Rays

Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted, and these were designated as alpha, beta, and gamma rays. The most penetrating nuclear radiation, the **gamma ray (γ ray)**, was later found to be an extremely high-frequency electromagnetic wave.

The lower end of the γ- ray frequency range overlaps the upper end of the X-ray range. Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. The name “gamma rays” is generally used for electromagnetic radiation emitted by a nucleus, while X-rays are generally produced by bombarding a target with energetic electrons in an X-ray tube. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Note:
Use this simulation to explore how light interacts with molecules in our atmosphere.
Explore how light interacts with molecules in our atmosphere.
Identify that absorption of light depends on the molecule and the type of light.
Relate the energy of the light to the resulting motion.
Identify that energy increases from microwave to ultraviolet.
Predict the motion of a molecule based on the type of light it absorbs.

Note:
Exercise:

Problem:

Check Your Understanding How do the electromagnetic waves for the different kinds of electromagnetic radiation differ?

Solution:

They fall into different ranges of wavelength, and therefore also different corresponding ranges of frequency.

Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v = f\lambda$, so that for electromagnetic waves, $c = f\lambda$, where f is the frequency, λ is the wavelength, and c is the speed of light.
- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.

Key Equations

Displacement current	$I_{\mathrm{d}} = \varepsilon_0 \frac{d\Phi_{\mathrm{E}}}{dt}$
Gauss’s law	$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = \frac{Q_{\mathrm{in}}}{\varepsilon_0}$
Gauss’s law for magnetism	$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}} = 0$
Faraday’s law	$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d\Phi_{\mathrm{m}}}{dt}$
Ampère-Maxwell law	$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I + \varepsilon_0 \mu_0 \frac{d\Phi_{\mathrm{E}}}{dt}$
Wave equation for plane EM wave	$\frac{\partial^2 E_y}{\partial x^2} = \varepsilon_0 \mu_0 \frac{\partial^2 E_y}{\partial t^2}$
Speed of EM waves	$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$
Ratio of *E* field to *B* field in electromagnetic wave	$c = \frac{E}{B}$
Energy flux (Poynting) vector	$\vec{\mathbf{S}} = \frac{1}{\mu_0} \vec{\mathbf{E}} \times \vec{\mathbf{B}}$
Average intensity of an electromagnetic wave	$I = S_{\mathrm{avg}} = \frac{c\varepsilon_0 E_0^2}{2} = \frac{cB_0^2}{2\mu_0} = \frac{E_0 B_0}{2\mu_0}$
Radiation pressure	$p = \begin{cases} I/c & \text{Perfect absorber} \\ 2I/c & \text{Perfect reflector} \end{cases}$

Conceptual Questions

Exercise:

Problem:

Compare the speed, wavelength, and frequency of radio waves and X-rays traveling in a vacuum.

Exercise:

Problem:

Accelerating electric charge emits electromagnetic radiation. How does this apply in each case: (a) radio waves, (b) infrared radiation.

Solution:

a. Radio waves are generally produced by alternating current in a wire or an oscillating electric field between two plates; b. Infrared radiation is commonly produced by heated bodies whose atoms and the charges in them vibrate at about the right frequency.

Exercise:

Problem:

Compare and contrast the meaning of the prefix “micro” in the names of SI units in the term *microwaves*.

Exercise:

Problem:

Part of the light passing through the air is scattered in all directions by the molecules comprising the atmosphere. The wavelengths of visible light are larger than molecular sizes, and the scattering is strongest for wavelengths of light closest to sizes of molecules.
(a) Which of the main colors of light is scattered the most? (b) Explain why this would give the sky its familiar background color at midday.

Solution:

a. blue; b. Light of longer wavelengths than blue passes through the air with less scattering, whereas more of the blue light is scattered in different directions in the sky to give it is blue color.

Exercise:

Problem:

When a bowl of soup is removed from a microwave oven, the soup is found to be steaming hot, whereas the bowl is only warm to the touch. Discuss the temperature changes that have occurred in terms of energy transfer.

Exercise:

Problem:

Certain orientations of a broadcast television antenna give better reception than others for a particular station. Explain.

Solution:

A typical antenna has a stronger response when the wires forming it are orientated parallel to the electric field of the radio wave.

Exercise:

Problem: What property of light corresponds to loudness in sound?

Exercise:

Problem: Is the visible region a major portion of the electromagnetic spectrum?

Solution:

No, it is very narrow and just a small portion of the overall electromagnetic spectrum.

Exercise:

Problem:

Can the human body detect electromagnetic radiation that is outside the visible region of the spectrum?

Exercise:

Problem:

Radio waves normally have their E and B fields in specific directions, whereas visible light usually has its E and B fields in random and rapidly changing directions that are perpendicular to each other and to the propagation direction. Can you explain why?

Solution:

Visible light is typically produced by changes of energies of electrons in randomly oriented atoms and molecules. Radio waves are typically emitted by an ac current flowing along a wire, that has fixed orientation and produces electric fields pointed in particular directions.

Exercise:

Problem: Give an example of resonance in the reception of electromagnetic waves.

Exercise:

Problem:

Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light).

Solution:

Radar can observe objects the size of an airplane and uses radio waves of about 0.5 cm in wavelength. Visible light can be used to view single biological cells and has wavelengths of about 10^{-7} m.

Exercise:

Problem: In which part of the electromagnetic spectrum are each of these waves: (a) f = 10.0 kHz, (b) $f = \lambda = 750\text{ nm}$, (c) $f = 1.25 \times 10^8$ Hz, (d) 0.30 nm

Exercise:

Problem:

In what range of electromagnetic radiation are the electromagnetic waves emitted by power lines in a country that uses 50-Hz ac current?

Solution:

ELF radio waves

Exercise:

Problem:

If a microwave oven could be modified to merely tune the waves generated to be in the infrared range instead of using microwaves, how would this affect the uneven heating of the oven?

Exercise:

Problem:

A leaky microwave oven in a home can sometimes cause interference with the homeowner’s WiFi system. Why?

Solution:

The frequency of 2.45 GHz of a microwave oven is close to the specific frequencies in the 2.4 GHz band used for WiFi.

Exercise:

Problem:

When a television news anchor in a studio speaks to a reporter in a distant country, there is sometimes a noticeable lag between when the anchor speaks in the studio and when the remote reporter hears it and replies. Explain what causes this delay.

Problems

Exercise:

Problem:

How many helium atoms, each with a radius of about 31 pm, must be placed end to end to have a length equal to one wavelength of 470 nm blue light?

Exercise:

Problem:

If you wish to detect details of the size of atoms (about 0.2 nm) with electromagnetic radiation, it must have a wavelength of about this size. (a) What is its frequency? (b) What type of electromagnetic radiation might this be?

Solution:

a. 1.5×10^{18} Hz; b. X-rays

Exercise:

Problem:

Find the frequency range of visible light, given that it encompasses wavelengths from 380 to 760 nm.

Exercise:

Problem:

(a) Calculate the wavelength range for AM radio given its frequency range is 540 to 1600 kHz. (b) Do the same for the FM frequency range of 88.0 to 108 MHz.

Solution:

a. The wavelength range is 187 m to 556 m. b. The wavelength range is 2.78 m to 3.41 m.

Exercise:

Problem:

Radio station WWVB, operated by the National Institute of Standards and Technology (NIST) from Fort Collins, Colorado, at a low frequency of 60 kHz, broadcasts a time synchronization signal whose range covers the entire continental US. The timing of the synchronization signal is controlled by a set of atomic clocks to an accuracy of 1×10^{-12} s, and repeats every 1 minute. The signal is used for devices, such as radio-controlled watches, that automatically synchronize with it at preset local times. WWVB's long wavelength signal tends to propagate close to the ground.
(a) Calculate the wavelength of the radio waves from WWVB.
(b) Estimate the error that the travel time of the signal causes in synchronizing a radio controlled watch in Norfolk, Virginia, which is 1570 mi (2527 km) from Fort Collins, Colorado.

Exercise:

Problem:

An outdoor WiFi unit for a picnic area has a 100-mW output and a range of about 30 m. What output power would reduce its range to 12 m for use with the same devices as before? Assume there are no obstacles in the way and that microwaves into the ground are simply absorbed.

Solution:

$P\prime = \left(\frac{12\,\mathrm{m}}{30\,\mathrm{m}}\right)^2 (100\,\mathrm{mW}) = 16\,\mathrm{mW}$

Exercise:

Problem:

7. The prefix “mega” (M) and “kilo” (k), when referring to amounts of computer data, refer to factors of 1024 or 2^{10} rather than 1000 for the prefix *kilo*, and $1024^2 = 2^{20}$ rather than 1,000,000 for the prefix *Mega* (M). If a wireless (WiFi) router transfers 150 Mbps of data, how many bits per second is that in decimal arithmetic?

Exercise:

Problem:

A computer user finds that his wireless router transmits data at a rate of 75 Mbps (megabits per second). Compare the average time to transmit one bit of data with the time difference between the wifi signal reaching an observer’s cell phone directly and by bouncing back to the observer from a wall 8.00 m past the observer.

Solution:

time for $1 \text{ bit} = 1.27 \times 10^{-8}$ s, difference in travel time is 5.34×10^{-8} s

Exercise:

Problem:

(a) The ideal size (most efficient) for a broadcast antenna with one end on the ground is one-fourth the wavelength $(\lambda/4)$ of the electromagnetic radiation being sent out. If a new radio station has such an antenna that is 50.0 m high, what frequency does it broadcast most efficiently? Is this in the AM or FM band? (b) Discuss the analogy of the fundamental resonant mode of an air column closed at one end to the resonance of currents on an antenna that is one-fourth their wavelength.

Exercise:

Problem:

What are the wavelengths of (a) X-rays of frequency 2.0×10^{17} Hz? (b) Yellow light of frequency 5.1×10^{14} Hz? (c) Gamma rays of frequency 1.0×10^{23} Hz?

Solution:

a. 1.5×10^{-9} m; b. 5.9×10^{-7} m; c. 3.0×10^{-15} m

Exercise:

Problem: For red light of $\lambda = 660$ nm, what are *f*, ω, and *k*?

Exercise:

Problem:

A radio transmitter broadcasts plane electromagnetic waves whose maximum electric field at a particular location is $1.55 \times 10^{-3}\ \text{V/m}$. What is the maximum magnitude of the oscillating magnetic field at that location? How does it compare with Earth's magnetic field?

Solution:

$5.17 \times 10^{-12}\ \text{T}$, the non-oscillating geomagnetic field of 25–65 μT is much larger

Exercise:

Problem:

(a) Two microwave frequencies authorized for use in microwave ovens are: 915 and 2450 MHz. Calculate the wavelength of each. (b) Which frequency would produce smaller hot spots in foods due to interference effects?

Exercise:

Problem:

During normal beating, the heart creates a maximum 4.00-mV potential across 0.300 m of a person's chest, creating a 1.00-Hz electromagnetic wave. (a) What is the maximum electric field strength created? (b) What is the corresponding maximum magnetic field strength in the electromagnetic wave? (c) What is the wavelength of the electromagnetic wave?

Solution:

a. $1.33 \times 10^{-2}\ \text{V/m}$; b. $4.44 \times 10^{-11}\ \text{T}$; c. $3.00 \times 10^{8}\ \text{m}$

Exercise:

Problem:

Distances in space are often quoted in units of light-years, the distance light travels in 1 year. (a) How many meters is a light-year? (b) How many meters is it to Andromeda, the nearest large galaxy, given that it is 2.54×10^{6} ly away? (c) The most distant galaxy yet discovered is 13.4×10^{9} ly away. How far is this in meters?

Exercise:

Problem:

A certain 60.0-Hz ac power line radiates an electromagnetic wave having a maximum electric field strength of 13.0 kV/m. (a) What is the wavelength of this very-low-frequency electromagnetic wave? (b) What type of electromagnetic radiation is this wave (b) What is its maximum magnetic field strength?

Solution:

a. $5.00 \times 10^{6} \mathrm{m}$; b. radio wave; c. $4.33 \times 10^{-5} \mathrm{T}$

Exercise:

Problem:

(a) What is the frequency of the 193-nm ultraviolet radiation used in laser eye surgery? (b) Assuming the accuracy with which this electromagnetic radiation can ablate (reshape) the cornea is directly proportional to wavelength, how much more accurate can this UV radiation be than the shortest visible wavelength of light?

Additional Problems

Exercise:

Problem:

In a region of space, the electric field is pointed along the *x*-axis, but its magnitude changes as described by

$E_x = (10\,\mathrm{N/C})\sin\,(20x - 500t)$

$E_y = E_z = 0$

where *t* is in nanoseconds and *x* is in cm. Find the displacement current through a circle of radius 3 cm in the $x = 0$ plane at $t = 0$.

Solution:

$I_\mathrm{d} = (10\,\mathrm{N/C})\left(8.845 \times 10^{-12}\,\mathrm{C^2/N \cdot m^2}\right)\pi(0.03\,\mathrm{m})^2\left(5000\frac{1}{\mathrm{s}}\right) = 1.25 \times 10^{-9}\,\mathrm{A}$

Exercise:

Problem:

A microwave oven uses electromagnetic waves of frequency $f = 2.45 \times 10^9\,\text{Hz}$ to heat foods. The waves reflect from the inside walls of the oven to produce an interference pattern of standing waves whose antinodes are hot spots that can leave observable pit marks in some foods. The pit marks are measured to be 6.0 cm apart. Use the method employed by Heinrich Hertz to calculate the speed of electromagnetic waves this implies.

Use the Appendix D for the next two exercises

Exercise:

Problem:

Galileo proposed measuring the speed of light by uncovering a lantern and having an assistant a known distance away uncover his lantern when he saw the light from Galileo’s lantern, and timing the delay. How far away must the assistant be for the delay to equal the human reaction time of about 0.25 s?

Solution:

$3.75 \times 10^7\,\text{km}$, which is much greater than Earth’s circumference

Exercise:

Problem:

Show that the wave equation in one dimension

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 f}{\partial t^2}$$

is satisfied by any doubly differentiable function of either the form $f\left(x - vt\right)$ or $f\left(x + vt\right)$.

Exercise:

Problem:

On its highest power setting, a microwave oven increases the temperature of 0.400 kg of spaghetti by $45.0\,^\circ\text{C}$ in 120 s. (a) What was the rate of energy absorption by the spaghetti, given that its specific heat is $3.76 \times 10^3\,\text{J/kg}\cdot{}^\circ\text{C}$? Assume the spaghetti is perfectly absorbing. (b) Find the average intensity of the microwaves, given that they are absorbed over a circular area 20.0 cm in diameter. (c) What is the peak electric field strength of the microwave? (d) What is its peak magnetic field strength?

Solution:

a. 564 W; b. $1.80 \times 10^4\ \text{W/m}^2$; c. $3.68 \times 10^3\ \text{V/m}$; d. $1.23 \times 10^{-5}\ \text{T}$

Exercise:

Problem:

A certain microwave oven projects 1.00 kW of microwaves onto a 30-cm-by-40-cm area. (a) What is its intensity in W/m^2 ? (b) Calculate the maximum electric field strength E_0 in these waves. (c) What is the maximum magnetic field strength B_0 ?

Exercise:

Problem:

Electromagnetic radiation from a 5.00-mW laser is concentrated on a 1.00-mm^2 area. (a) What is the intensity in W/m^2 ? (b) Suppose a 2.00-nC electric charge is in the beam. What is the maximum electric force it experiences? (c) If the electric charge moves at 400 m/s, what maximum magnetic force can it feel?

Solution:

a. $5.00 \times 10^3\ \text{W/m}^2$; b. 3.88×10^{-6} N; c. 5.18×10^{-12} N

Exercise:

Problem:

A 200-turn flat coil of wire 30.0 cm in diameter acts as an antenna for FM radio at a frequency of 100 MHz. The magnetic field of the incoming electromagnetic wave is perpendicular to the coil and has a maximum strength of $1.00 \times 10^{-12}\ \text{T}$. (a) What power is incident on the coil? (b) What average emf is induced in the coil over one-fourth of a cycle? (c) If the radio receiver has an inductance of $2.50\ \mu\text{H}$, what capacitance must it have to resonate at 100 MHz?

Exercise:

Problem:

Suppose a source of electromagnetic waves radiates uniformly in all directions in empty space where there are no absorption or interference effects. (a) Show that the intensity is inversely proportional to r^2, the distance from the source squared. (b) Show that the magnitudes of the electric and magnetic fields are inversely proportional to r.

Solution:

a. $I = \frac{P}{A} = \frac{P}{4\pi r^2} \propto \frac{1}{r^2}$; b. $I \propto E_0^2\ ,\ B_0^2 \Rightarrow E_0^2\ ,\ B_0^2 \propto \frac{1}{r^2} \Rightarrow E_0, B_0 \propto \frac{1}{r}$

Exercise:

Problem:

A radio station broadcasts its radio waves with a power of 50,000 W. What would be the intensity of this signal if it is received on a planet orbiting Proxima Centuri, the closest star to our Sun, at 4.243 ly away?

Exercise:

Problem:

The Poynting vector describes a flow of energy whenever electric and magnetic fields are present. Consider a long cylindrical wire of radius *r* with a current *I* in the wire, with resistance *R* and voltage *V*. From the expressions for the electric field along the wire and the magnetic field around the wire, obtain the magnitude and direction of the Poynting vector at the surface. Show that it accounts for an energy flow into the wire from the fields around it that accounts for the Ohmic heating of the wire.

Solution:

$$\begin{aligned}\text{Power into the wire} &= \int \vec{\mathbf{S}} \cdot d\vec{\mathbf{A}} = \left(\frac{1}{\mu_0} EB\right)(2\pi r L) \\ &= \frac{1}{\mu_0}\left(\frac{V}{L}\right)\left(\frac{\mu_0 i}{2\pi r}\right)(2\pi r L) = iV = i^2 R\end{aligned}$$

Exercise:

Problem:

The Sun’s energy strikes Earth at an intensity of $1.37\ \text{kW/m}^2$. Assume as a model approximation that all of the light is absorbed. (Actually, about 30% of the light intensity is reflected out into space.)
(a) Calculate the total force that the Sun’s radiation exerts on Earth.
(b) Compare this to the force of gravity between the Sun and Earth.
Earth’s mass is 5.972×10^{24} kg.

Exercise:

Problem:

If a *Lightsail* spacecraft were sent on a Mars mission, by what ratio of the final force to the initial force would its propulsion be reduced when it reached Mars?

Solution:

0.431

Exercise:

Problem:

Lunar astronauts placed a reflector on the Moon’s surface, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. (a) To what accuracy in meters can the distance to the Moon be determined, if this time can be measured to 0.100 ns? (b) What percent accuracy is this, given the average distance to the Moon is 384,400 km?

Exercise:

Problem:

Radar is used to determine distances to various objects by measuring the round-trip time for an echo from the object. (a) How far away is the planet Venus if the echo time is 1000 s? (b) What is the echo time for a car 75.0 m from a highway police radar unit? (c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?

Solution:

a. 1.5×10^{11}m; b. 5.0×10^{-7}s; c. 33 ns

Exercise:

Problem:

Calculate the ratio of the highest to lowest frequencies of electromagnetic waves the eye can see, given the wavelength range of visible light is from 380 to 760 nm. (Note that the ratio of highest to lowest frequencies the ear can hear is 1000.)

Exercise:

Problem:

How does the wavelength of radio waves for an AM radio station broadcasting at 1030 KHz compare with the wavelength of the lowest audible sound waves (of 20 Hz). The speed of sound in air at $20\ ^\circ\mathrm{C}$ is about 343 m/s.

Solution:

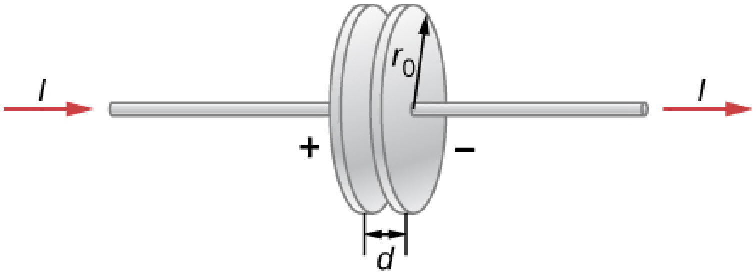
$$\begin{aligned}\text{sound: } \lambda_{\text{sound}} &= \frac{v_s}{f} = \frac{343\ \mathrm{m/s}}{20.0\ \mathrm{Hz}} = 17.2\ \mathrm{m} \\ \text{radio: } \lambda_{\text{radio}} &= \frac{c}{f} = \frac{3.00 \times 10^8\ \mathrm{m/s}}{1030 \times 10^3\ \mathrm{Hz}} = 291\ \mathrm{m}\text{; or } 17.1\ \lambda_{\text{sound}}\end{aligned}$$

Challenge Problems

Exercise:

Problem:

A parallel-plate capacitor with plate separation d is connected to a source of emf that places a time-dependent voltage $V(t)$ across its circular plates of radius r_0 and area $A = \pi r_0^2$ (see below).



(a) Write an expression for the time rate of change of energy inside the capacitor in terms of $V(t)$ and $dV(t)/dt$.

(b) Assuming that $V(t)$ is increasing with time, identify the directions of the electric field lines inside the capacitor and of the magnetic field lines at the edge of the region between the plates, and then the direction of the Poynting vector $\vec{\mathbf{S}}$ at this location.

(c) Obtain expressions for the time dependence of $E(t)$, for $B(t)$ from the displacement current, and for the magnitude of the Poynting vector at the edge of the region between the plates.

(d) From $\vec{\mathbf{S}}$, obtain an expression in terms of $V(t)$ and $dV(t)/dt$ for the rate at which electromagnetic field energy enters the region between the plates.

(e) Compare the results of parts (a) and (d) and explain the relationship between them.

Exercise:

Problem:

A particle of cosmic dust has a density $\rho = 2.0\,\text{g}/\text{cm}^3$. (a) Assuming the dust particles are spherical and light absorbing, and are at the same distance as Earth from the Sun, determine the particle size for which radiation pressure from sunlight is equal to the Sun’s force of gravity on the dust particle. (b) Explain how the forces compare if the particle radius is smaller. (c) Explain what this implies about the sizes of dust particle likely to be present in the inner solar system compared with outside the Oort cloud.

Solution:

a. $0.29\ \mu\text{m}$; b. The radiation pressure is greater than the Sun’s gravity if the particle size is smaller, because the gravitational force varies as the radius cubed while the radiation pressure varies as the radius squared. c. The radiation force outward implies that particles smaller than this are less likely to be near the Sun than outside the range of the Sun’s radiation pressure.

Glossary

gamma ray (γ ray)

extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons; the lower end of the γ -ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

infrared radiation

region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from $0.74\ \mu\text{m to } 300\ \mu\text{m}$

microwaves

electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

radar

common application of microwaves; radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

radio waves

electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

thermal agitation

thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

ultraviolet radiation

electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm visible light

narrow segment of the electromagnetic spectrum to which the normal human eye responds, from about 400 to 750 nm

X-ray

invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

Introduction class="introduction"

Due to total internal reflection, an underwater swimmer’s image is reflected back into the water where the camera is located. The circular ripple in the image center is actually on the water surface. Due to the viewing angle, total internal reflection is not occurring at the top edge of this image, and we can see a view of activities on the pool deck. (credit: modification of work by “jayhem”/Flickr)

Our investigation of light revolves around two questions of fundamental importance: (1) What is the nature of light, and (2) how does light behave under various circumstances? Answers to these questions can be found in Maxwell's equations (in Electromagnetic Waves), which predict the existence of electromagnetic waves and their behavior. Examples of light include radio and infrared waves, visible light, ultraviolet radiation, and X-rays. Interestingly, not all light phenomena can be explained by Maxwell's theory. Experiments performed early in the twentieth century showed that light has corpuscular, or particle-like, properties. The idea that light can display both wave and particle characteristics is called *wave-particle duality*, which is examined in Photons and Matter Waves.

In this chapter, we study the basic properties of light. In the next few chapters, we investigate the behavior of light when it interacts with optical devices such as mirrors, lenses, and apertures.

The Propagation of Light

By the end of this section, you will be able to:

- Determine the index of refraction, given the speed of light in a medium
- List the ways in which light travels from a source to another location

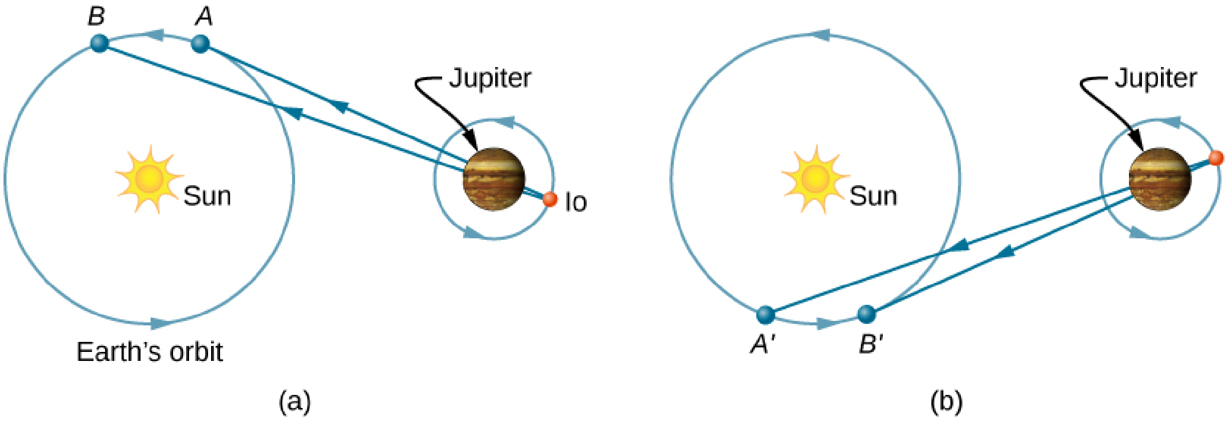
The speed of light in a vacuum c is one of the fundamental constants of physics. As you will see when you reach Relativity, it is a central concept in Einstein's theory of relativity. As the accuracy of the measurements of the speed of light improved, it was found that different observers, even those moving at large velocities with respect to each other, measure the same value for the speed of light. However, the speed of light does vary in a precise manner with the material it traverses. These facts have far-reaching implications, as we will see in later chapters.

The Speed of Light: Early Measurements

The first measurement of the speed of light was made by the Danish astronomer Ole Roemer (1644–1710) in 1675. He studied the orbit of Io, one of the four large moons of Jupiter, and found that it had a period of revolution of 42.5 h around Jupiter. He also discovered that this value fluctuated by a few seconds, depending on the position of Earth in its orbit around the Sun. Roemer realized that this fluctuation was due to the finite speed of light and could be used to determine c.

Roemer found the period of revolution of Io by measuring the time interval between successive eclipses by Jupiter. [link](a) shows the planetary configurations when such a measurement is made from Earth in the part of its orbit where it is receding from Jupiter. When Earth is at point A, Earth, Jupiter, and Io are aligned. The next time this alignment occurs, Earth is at point B, and the light carrying that information to Earth must travel to that point. Since B is farther from Jupiter than A, light takes more time to reach Earth when Earth is at B. Now imagine it is about 6 months later, and the planets are arranged as in part (b) of the figure. The measurement of Io's period begins with Earth at point A' and Io eclipsed by Jupiter. The next eclipse then occurs when Earth is at point B', to which the light carrying the information of this eclipse must travel. Since B' is closer to Jupiter than

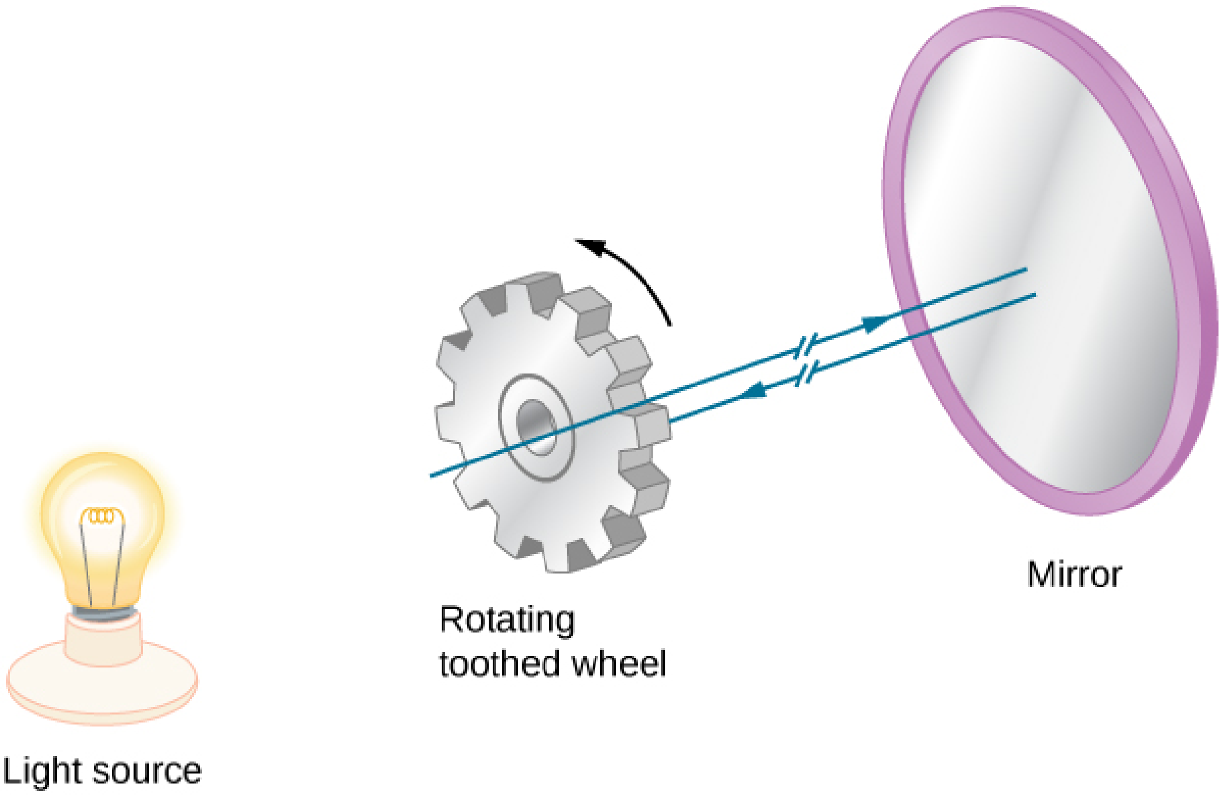
A', light takes less time to reach Earth when it is at B'. This time interval between the successive eclipses of Io seen at A' and B' is therefore less than the time interval between the eclipses seen at A and B. By measuring the difference in these time intervals and with appropriate knowledge of the distance between Jupiter and Earth, Roemer calculated that the speed of light was $2.0\;\times\;10^8\;\text{m/s}$, which is 33% below the value accepted today.



(a) (b)

Roemer's astronomical method for determining the speed of light. Measurements of Io's period done with the configurations of parts (a) and (b) differ, because the light path length and associated travel time increase from A to B (a) but decrease from A' to B' (b).

The first successful terrestrial measurement of the speed of light was made by Armand Fizeau (1819–1896) in 1849. He placed a toothed wheel that could be rotated very rapidly on one hilltop and a mirror on a second hilltop 8 km away ([link]). An intense light source was placed behind the wheel, so that when the wheel rotated, it chopped the light beam into a succession of pulses. The speed of the wheel was then adjusted until no light returned to the observer located behind the wheel. This could only happen if the wheel rotated through an angle corresponding to a displacement of $(n+½)$ teeth, while the pulses traveled down to the mirror and back. Knowing the rotational speed of the wheel, the number of teeth on the wheel, and the distance to the mirror, Fizeau determined the speed of light to be $3.15 \times 10^8 \text{ m/s}$, which is only 5% too high.



Fizeau's method for measuring the speed of light. The teeth of the wheel block the reflected light upon return when the wheel is rotated at a rate that matches the light travel time to and from the mirror.

The French physicist Jean Bernard Léon Foucault (1819–1868) modified Fizeau's apparatus by replacing the toothed wheel with a rotating mirror. In 1862, he measured the speed of light to be $2.98 \times 10^8 \text{ m/s}$, which is within 0.6% of the presently accepted value. Albert Michelson (1852–1931) also used Foucault's method on several occasions to measure the speed of light. His first experiments were performed in 1878; by 1926, he had refined the technique so well that he found *c* to be $(2.99796 \pm 4) \times 10^8 \text{ m/s}$.

Today, the speed of light is known to great precision. In fact, the speed of light in a vacuum *c* is so important that it is accepted as one of the basic physical quantities and has the value

Note:
Equation:

$$c = 2.99792458\ \times\ 10^8\,\mathrm{m/s} \approx 3.00\ \times\ 10^8\,\mathrm{m/s}$$

where the approximate value of $3.00\ \times\ 10^8\,\mathrm{m/s}$ is used whenever three-digit accuracy is sufficient.

Speed of Light in Matter

The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction varies with different atoms, crystal lattices, and other substructures. We can define a constant of a material that describes the speed of light in it, called the **index of refraction** *n*:

Note:
Equation:

$$n = \frac{c}{v}$$

where *v* is the observed speed of light in the material.

Since the speed of light is always less than *c* in matter and equals *c* only in a vacuum, the index of refraction is always greater than or equal to one; that is, $n \geq 1$. [link] gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors separated by a prism, as we will see in Dispersion.) Note that for gases, *n* is close to 1.0. This seems reasonable, since atoms in gases are widely separated, and light travels at *c* in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light *v* in a medium varies considerably from its value *c* in a vacuum, it is still a large speed.

Medium	***n***
Gases at 0°C, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20°C	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461

Medium	***n***
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice (at 0°C)	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Index of Refraction in Various MediaFor light with a wavelength of 589 nm in a vacuum

Example:

Speed of Light in Jewelry

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

We can calculate the speed of light in a material *v* from the index of refraction *n* of the material, using the equation $n = c/v$.

Solution

Rearranging the equation $n = c/v$ for *v* gives us

Equation:

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in [link], and *c* is given in [link]. Entering these values in the equation gives

Equation:

$$v = \frac{3.00 \times 10^8 \, \text{m/s}}{1.923} = 1.56 \times 10^8 \, \text{m/s}.$$

Significance

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in [link] that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Note:

Exercise:

Problem:

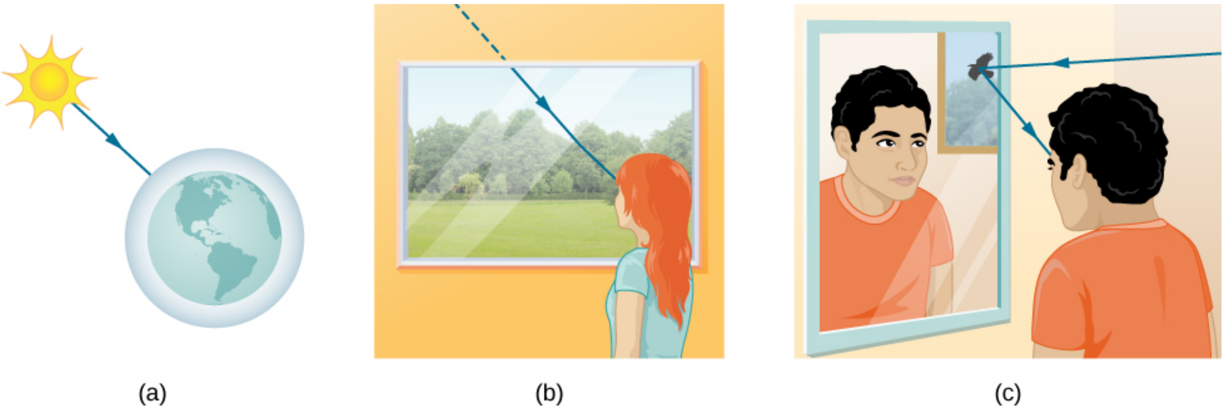
Check Your Understanding [link] shows that ethanol and fresh water have very similar indices of refraction. By what percentage do the speeds of light in these liquids differ?

Solution:

2.1% (to two significant figures)

The Ray Model of Light

You have already studied some of the wave characteristics of light in the previous chapter on Electromagnetic Waves. In this chapter, we start mainly with the ray characteristics. There are three ways in which light can travel from a source to another location ([link]). It can come directly from the source through empty space, such as from the Sun to Earth. Or light can travel through various media, such as air and glass, to the observer. Light can also arrive after being reflected, such as by a mirror. In all of these cases, we can model the path of light as a straight line called a **ray**.



Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth, traveling through empty space directly from the source. (b) Light can reach a person by traveling through media like air and glass. (c) Light can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Experiments show that when light interacts with an object several times larger than its wavelength, it travels in straight lines and acts like a ray. Its wave characteristics are not pronounced in such situations. Since the wavelength of visible light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when visible light encounters anything large enough that we can observe it with unaided eyes, such as a coin, it acts like a ray, with generally negligible wave characteristics.

In all of these cases, we can model the path of light as straight lines. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray. The word “ray” comes from mathematics and here means a straight line that originates at some point. It is acceptable to visualize light rays as laser rays. The *ray model of light* describes the path of light as straight lines.

Since light moves in straight lines, changing directions when it interacts with materials, its path is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called **geometric optics**. Two laws govern how light changes direction when it interacts with matter. These are the *law of reflection*, for situations in which light bounces off matter, and the *law of refraction*, for situations in which light passes through matter. We will examine more about each of these laws in upcoming sections of this chapter.

Summary

- The speed of light in a vacuum is $c = 2.99792458\ \times\ 10^8\,\text{m/s} \approx 3.00\ \times\ 10^8\,\text{m/s}.$
- The index of refraction of a material is $n = c/v$, where v is the speed of light in a material and c is the speed of light in a vacuum.
- The ray model of light describes the path of light as straight lines. The part of optics dealing with the ray aspect of light is called geometric optics.

- Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; and (3) after being reflected from a mirror.

Conceptual Questions

Exercise:

Problem:

Under what conditions can light be modeled like a ray? Like a wave?

Solution:

Light can be modeled as a ray when devices are large compared to wavelength, and as a wave when devices are comparable or small compared to wavelength.

Exercise:

Problem:

Why is the index of refraction always greater than or equal to 1?

Exercise:

Problem:

Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.

Solution:

This fact simply proves that the speed of light is greater than that of sound. If one knows the distance to the location of the lightning and the speed of sound, one could, in principle, determine the speed of light from the data. In practice, because the speed of light is so great, the data would have to be known to impractically high precision.

Exercise:

Problem:

Speculate as to what physical process might be responsible for light traveling more slowly in a medium than in a vacuum.

Problems

Exercise:

Problem: What is the speed of light in water? In glycerine?

Exercise:

Problem: What is the speed of light in air? In crown glass?

Solution:

$2.99705 \times 10^8 \text{ m/s}$; $1.97 \times 10^8 \text{ m/s}$

Exercise:

Problem:

Calculate the index of refraction for a medium in which the speed of light is $2.012 \times 10^8 \text{ m/s}$, and identify the most likely substance based on [link].

Exercise:

Problem:

In what substance in [link] is the speed of light $2.290 \times 10^8 \text{ m/s}$?

Solution:

ice at 0°C

Exercise:

Problem:

There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is $3.84 \times 10^5 \text{ km}$ away, would the light first arrive on Earth?

Exercise:

Problem:

Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

Solution:

1.03 ns

Exercise:

Problem:

Compare the time it takes for light to travel 1000 m on the surface of Earth and in outer space.

Exercise:

Problem:

How far does light travel underwater during a time interval of 1.50×10^{-6} s?

Solution:

337 m

Glossary

geometric optics
part of optics dealing with the ray aspect of light

index of refraction
for a material, the ratio of the speed of light in a vacuum to that in a material

ray
straight line that originates at some point

The Law of Reflection

By the end of this section, you will be able to:

- Explain the reflection of light from polished and rough surfaces
- Describe the principle and applications of corner reflectors

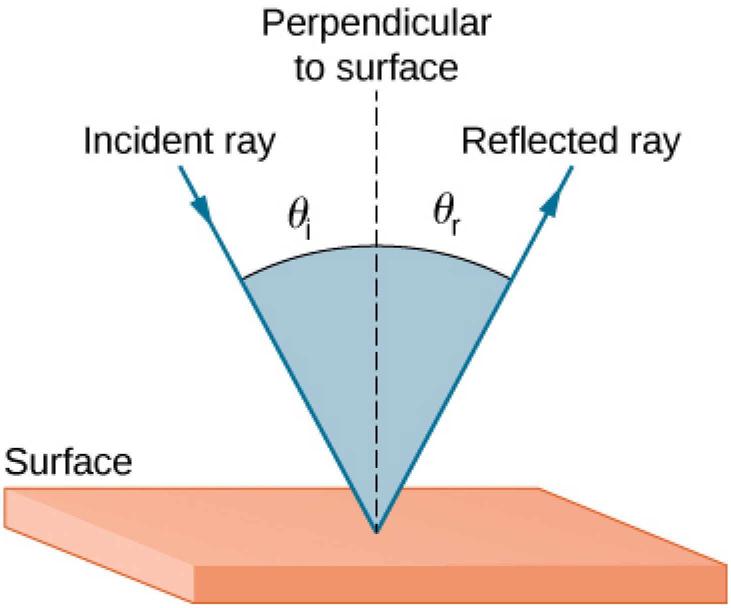
Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at a piece of white paper, you are seeing light scattered from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The **law of reflection** states that the angle of reflection equals the angle of incidence, or

Note:
Equation:

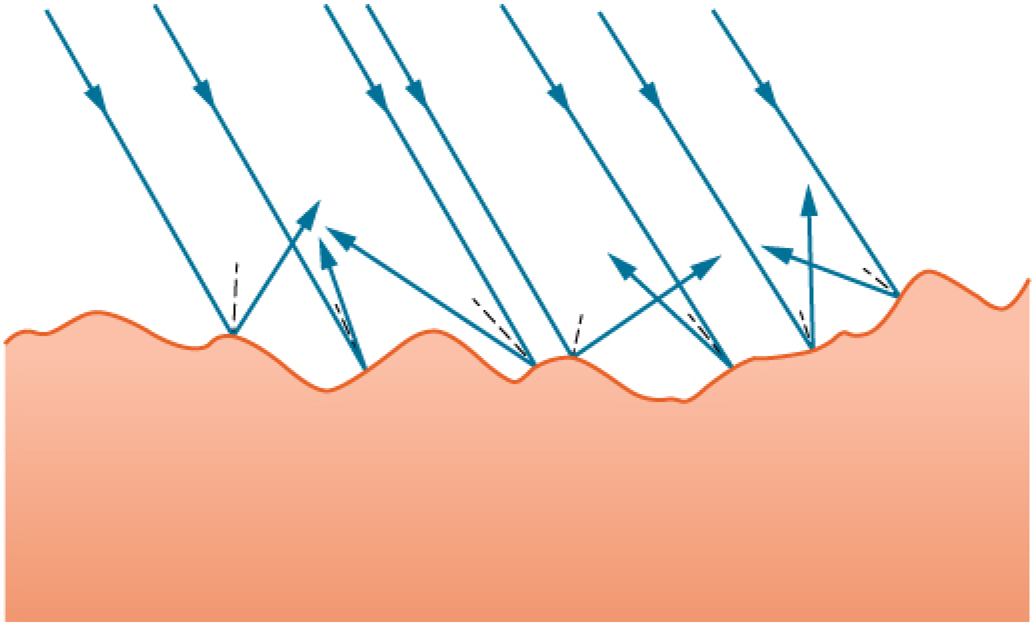
$$\theta_{\mathrm{r}} = \theta_{\mathrm{i}}$$

The law of reflection is illustrated in [link], which also shows how the angle of incidence and angle of reflection are measured relative to the perpendicular to the surface at the point where the light ray strikes.

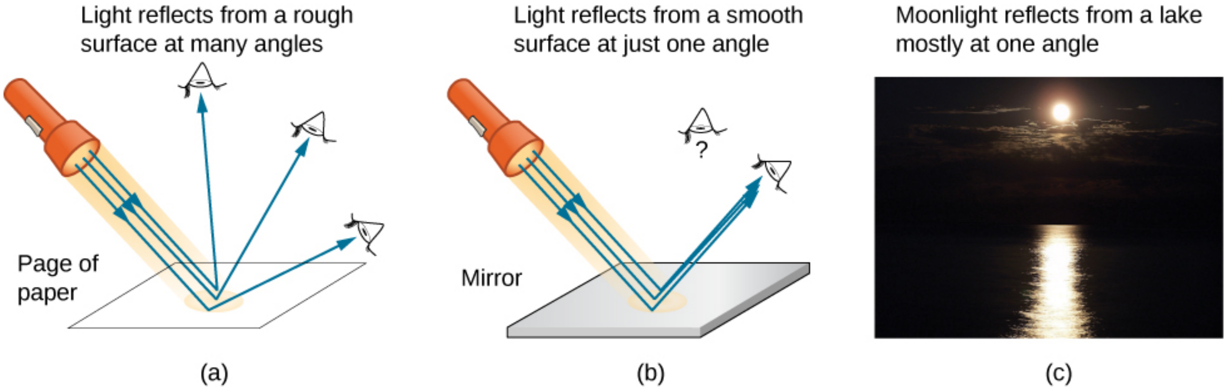


The law of reflection states that the angle of reflection equals the angle of incidence—$\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

We expect to see reflections from smooth surfaces, but [link] illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as shown in [link](a). People, clothing, leaves, and walls all have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in [link](b). When the Moon reflects from a lake, as shown in [link](c), a combination of these effects takes place.

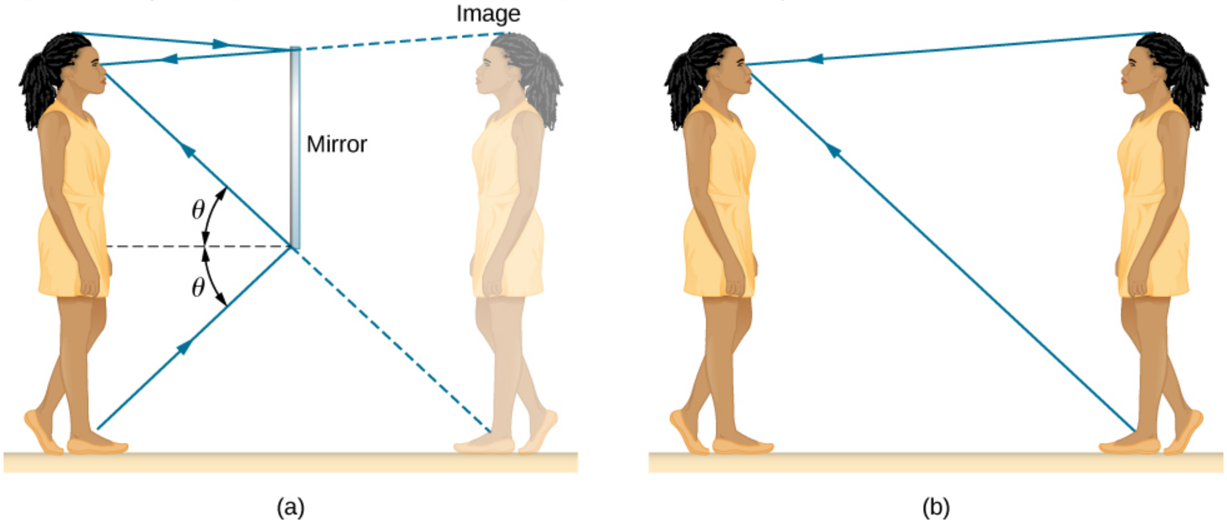


Light is diffused when it reflects from a rough surface. Here, many parallel rays are incident, but they are reflected at many different angles, because the surface is rough.


Light reflects from a rough surface at many angles
Page of paper
(a)
Light reflects from a smooth surface at just one angle
?
Mirror
(b)
Moonlight reflects from a lake mostly at one angle
(c)

(a) When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light. (b) A mirror illuminated by many parallel rays reflects them in only one direction, because its surface is very smooth. Only the observer at a particular angle sees the reflected light. (c) Moonlight is spread out when it is reflected by the lake, because the surface is shiny but uneven. (credit c: modification of work by Diego Torres Silvestre)

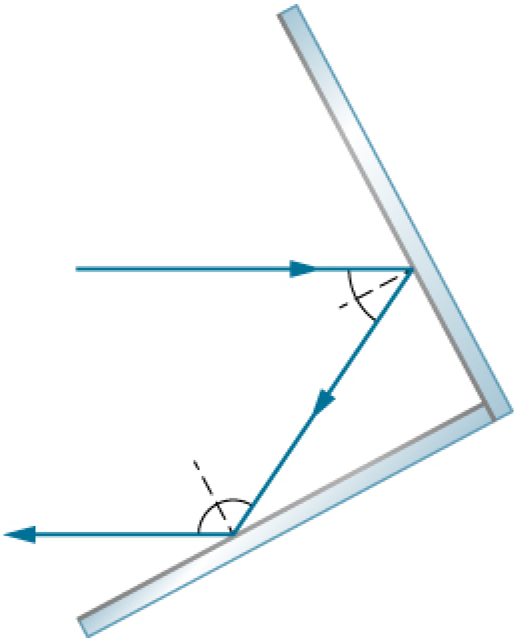
When you see yourself in a mirror, it appears that the image is actually behind the mirror ([link]). We see the light coming from a direction determined by the law of reflection. The angles are such that the image is exactly the same distance behind the mirror as you stand in front of the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of your imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (which are optical instruments themselves). The precise manner in which images are formed by mirrors and lenses is discussed in an upcoming chapter on Geometric Optics and Image Formation.



(a) Your image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be behind the mirror at the same distance away as (b) if you were looking at your twin directly, with no mirror.

Corner Reflectors (Retroreflectors)

A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came ([link]). This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. (For proof, see [link] at the end of this section.) Such an object is called a **corner reflector**, since the light bounces from its inside corner. Corner reflectors are a subclass of retroreflectors, which all reflect rays back in the directions from which they came. Although the geometry of the proof is much more complex, corner reflectors can also be built with three mutually perpendicular reflecting surfaces and are useful in three-dimensional applications.



A light ray that strikes two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came.

Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. Rather than simply reflecting light over a wide angle, retroreflection ensures high visibility if the observer and the light source are located together, such as a car’s driver and headlights. The Apollo astronauts placed a true corner reflector on the Moon ([link]). Laser signals from Earth can be bounced from that corner reflector to measure the gradually increasing distance to the Moon of a few centimeters per year.

(a) Astronauts placed a corner reflector on the Moon to measure its gradually increasing orbital distance. (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit a: modification of work by NASA; credit b: modification of work by “Julo”/Wikimedia Commons)

Working on the same principle as these optical reflectors, corner reflectors are routinely used as radar reflectors ([link]) for radio-frequency applications. Under most circumstances, small boats made of fiberglass or wood do not strongly reflect radio waves emitted by radar systems. To make these boats visible to radar (to avoid collisions, for example), radar reflectors are attached to boats, usually in high places.

A radar reflector hoisted on a sailboat is a type of corner reflector. (credit: Tim Sheerman-Chase)

As a counterexample, if you are interested in building a stealth airplane, radar reflections should be minimized to evade detection. One of the design considerations would then be to avoid building 90° corners into the airframe.

Summary

- When a light ray strikes a smooth surface, the angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.

Conceptual Questions

Exercise:

Problem:

Using the law of reflection, explain how powder takes the shine off of a person’s nose. What is the name of the optical effect?

Solution:

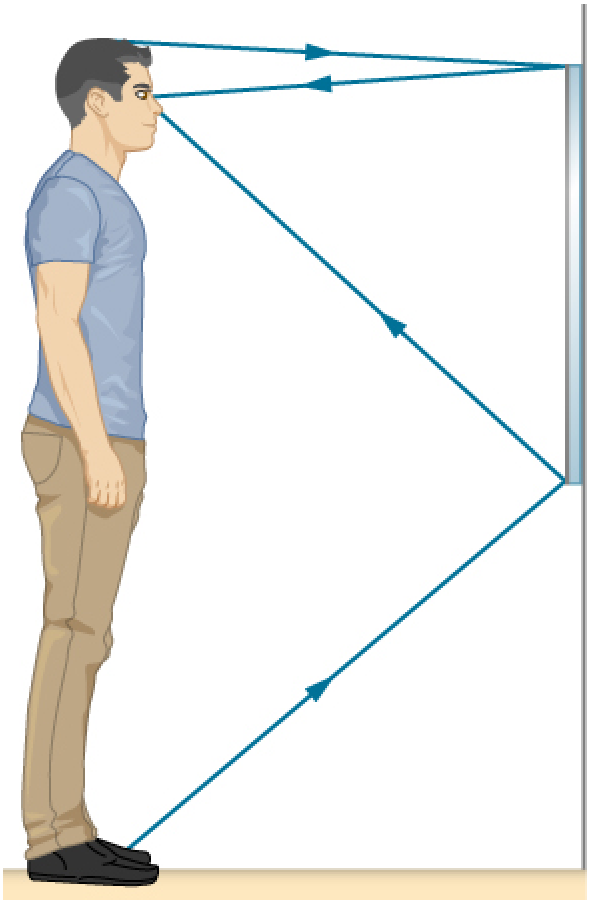
Powder consists of many small particles with randomly oriented surfaces. This leads to diffuse reflection, reducing shine.

Problems

Exercise:

Problem:

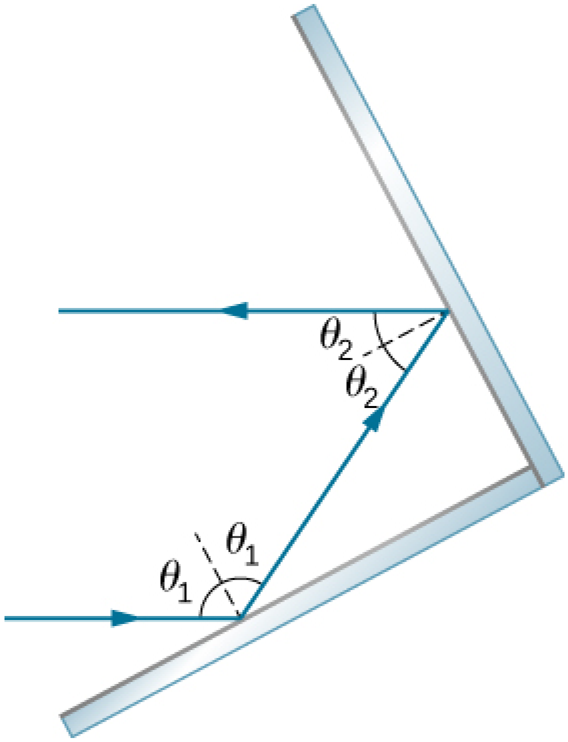
Suppose a man stands in front of a mirror as shown below. His eyes are 1.65 m above the floor and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man’s height?



Exercise:

Problem:

Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated below.



Solution:

proof

Exercise:

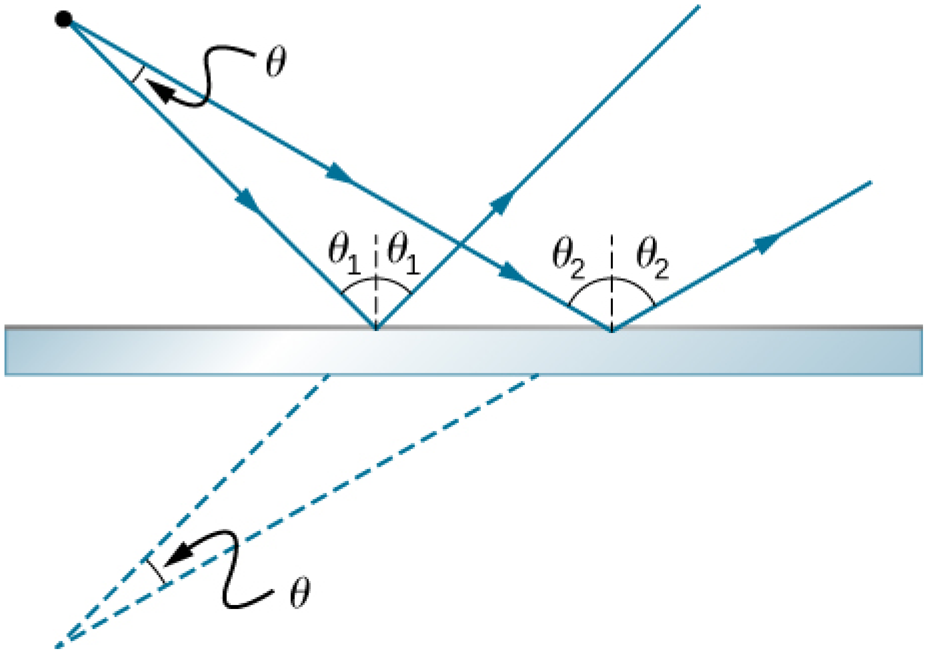
Problem:

On the Moon’s surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth’s atmosphere? Assume the distance to the Moon is precisely $3.84\ \times\ 10^8\ \text{m}$ and Earth’s atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

Exercise:

Problem:

A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ (see below). Show that after striking a plane mirror, the angle between their directions remains θ.



Solution:

proof

Glossary

corner reflector
object consisting of two (or three) mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

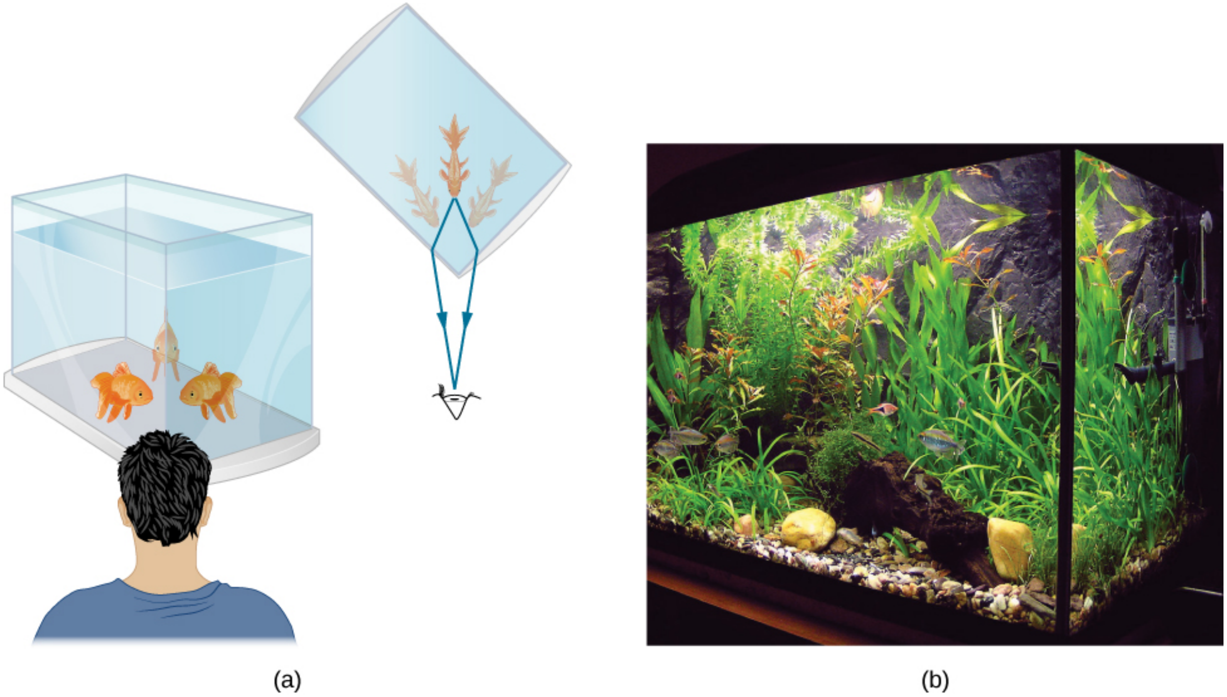
law of reflection
angle of reflection equals the angle of incidence

Refraction

By the end of this section, you will be able to:

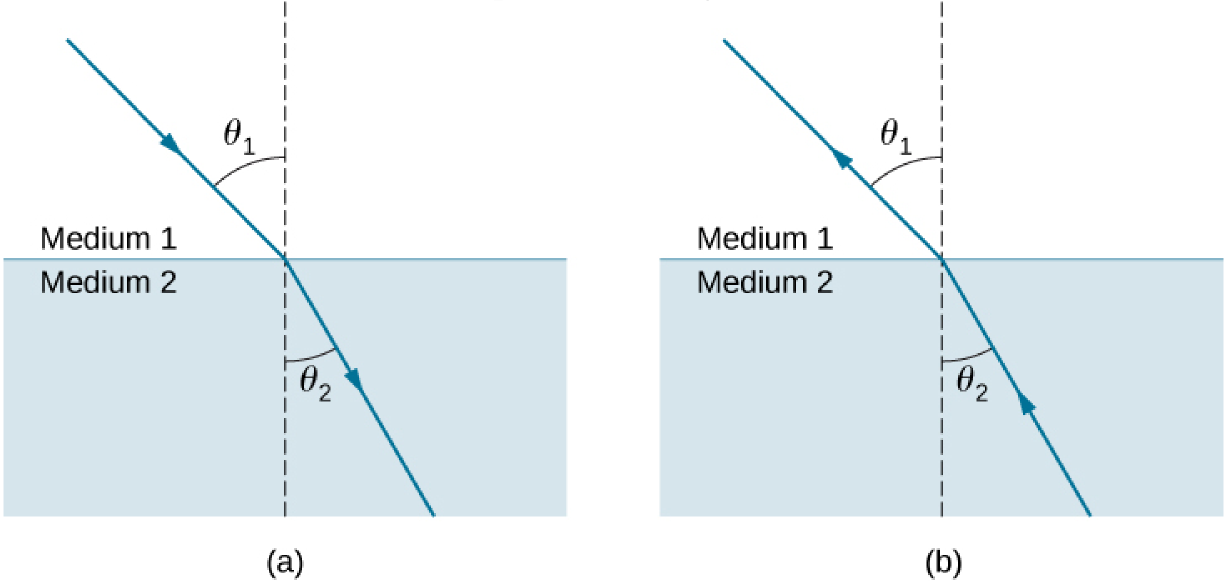
- Describe how rays change direction upon entering a medium
- Apply the law of refraction in problem solving

You may often notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places ([link]). This happens because light coming from the fish to you changes direction when it leaves the tank, and in this case, it can travel two different paths to get to your eyes. The changing of a light ray’s direction (loosely called bending) when it passes through substances of different refractive indices is called **refraction** and is related to changes in the speed of light, $v = c/n$. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to data transmission through optical fibers.



(a) Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena. (b) This image shows refraction of light from a fish near the top of a fish tank.

[link] shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light is reflected from the surface, but for now we concentrate on the light that is transmitted.) The change in direction of the light ray depends on the relative values of the indices of refraction (The Propagation of Light) of the two media involved. In the situations shown, medium 2 has a greater index of refraction than medium 1. Note that as shown in [link](a), the direction of the ray moves closer to the perpendicular when it progresses from a medium with a lower index of refraction to one with a higher index of refraction. Conversely, as shown in [link](b), the direction of the ray moves away from the perpendicular when it progresses from a medium with a higher index of refraction to one with a lower index of refraction. The path is exactly reversible.



The change in direction of a light ray depends on how the index of refraction changes when it crosses from one medium to another. In the situations shown here, the index of refraction is greater in medium 2 than in medium 1. (a) A ray of light moves closer to the perpendicular when entering a medium with a higher index of refraction. (b) A ray of light moves away from the perpendicular when entering a medium with a lower index of refraction.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction and thus a large change in angle. The exact mathematical relationship is the **law of refraction**, or Snell’s law, after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. The law of refraction is stated in equation form as

Note:
Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

Here n_1 and n_2 are the indices of refraction for media 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in media 1 and 2. The incoming ray is called the incident ray, the outgoing ray is called the refracted ray, and the associated angles are the incident angle and the refracted angle, respectively.

Snell’s experiments showed that the law of refraction is obeyed and that a characteristic index of refraction n could be assigned to a given medium and its value measured. Snell was not aware that the speed of light varied in different media, a key fact used when we derive the law of refraction theoretically using Huygens’s principle in Huygens’s Principle.

Example:

Determining the Index of Refraction

Find the index of refraction for medium 2 in [link](a), assuming medium 1 is air and given that the incident angle is 30.0° and the angle of refraction is 22.0°.

Strategy

The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus, $n_1 = 1.00$ here. From the given information, $\theta_1 = 30.0^\circ$ and $\theta_2 = 22.0^\circ$. With this information, the only unknown in Snell's law is n_2, so we can use Snell's law to find it.

Solution

From Snell's law we have

Equation:

$$\begin{array}{rcl} n_1 \sin\theta_1 & = & n_2 \sin\theta_2 \\ n_2 & = & n_1 \frac{\sin\theta_1}{\sin\theta_2}. \end{array}$$

Entering known values,

Equation:

$$n_2 = 1.00 \frac{\sin 30.0^\circ}{\sin 22.0^\circ} = \frac{0.500}{0.375} = 1.33.$$

Significance

This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today, we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Note:

Explore bending of light between two media with different indices of refraction. Use the "Intro" simulation and see how changing from air to water to glass changes the bending angle. Use the protractor tool to measure the angles and see if you can recreate the configuration in [link]. Also by measurement, confirm that the angle of reflection equals the angle of incidence.

Example:

A Larger Change in Direction

Suppose that in a situation like that in [link], light goes from air to diamond and that the incident angle is 30.0°. Calculate the angle of refraction θ_2 in the diamond.

Strategy

Again, the index of refraction for air is taken to be $n_1 = 1.00$, and we are given $\theta_1 = 30.0^\circ$. We can look up the index of refraction for diamond in [link], finding $n_2 = 2.419$. The only unknown in Snell's law is θ_2, which we wish to determine.

Solution

Solving Snell's law for $\sin\theta_2$ yields

Equation:

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1.$$

Entering known values,

Equation:

$$\sin\theta_2 = \frac{1.00}{2.419}\sin 30.0^\circ = (0.413)(0.500) = 0.207.$$

The angle is thus

Equation:

$$\theta_2 = \sin^{-1}(0.207) = 11.9^\circ.$$

Significance

For the same 30.0° angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9° rather than 22.0°—see [link]). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

Note:
Exercise:

Problem:

Check Your Understanding In [link], the solid with the next highest index of refraction after diamond is zircon. If the diamond in [link] were replaced with a piece of zircon, what would be the new angle of refraction?

Solution:

15.1°

Summary

- The change of a light ray’s direction when it passes through variations in matter is called refraction.
- The law of refraction, also called Snell’s law, relates the indices of refraction for two media at an interface to the change in angle of a light ray passing through that interface.

Conceptual Questions

Exercise:

Problem:

Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.

Exercise:

Problem:

Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?

Solution:

“toward” when increasing *n* (air to water, water to glass); “away” when decreasing *n* (glass to air)

Exercise:

Problem:

Explain why an object in water always appears to be at a depth shallower than it actually is?

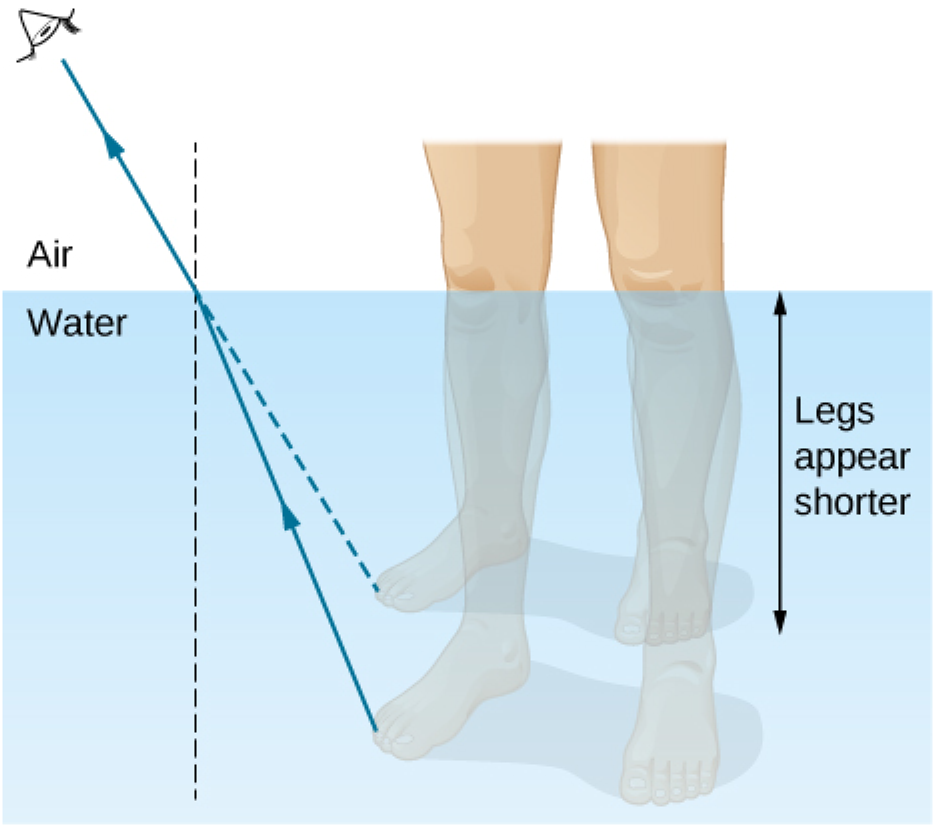
Exercise:

Problem:

Explain why a person’s legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.

Solution:

A ray from a leg emerges from water after refraction. The observer in air perceives an apparent location for the source, as if a ray traveled in a straight line. See the dashed ray below.



Exercise:

Problem:

Explain why an oar that is partially submerged in water appears bent.

Problems

Unless otherwise specified, for problems 1 through 10, the indices of refraction of glass and water should be taken to be 1.50 and 1.333, respectively.

Exercise:

Problem:

A light beam in air has an angle of incidence of $35°$ at the surface of a glass plate. What are the angles of reflection and refraction?

Exercise:

Problem:

A light beam in air is incident on the surface of a pond, making an angle of 20° with respect to the surface. What are the angles of reflection and refraction?

Solution:

reflection, 70°; refraction, 45°

Exercise:

Problem:

When a light ray crosses from water into glass, it emerges at an angle of 30° with respect to the normal of the interface. What is its angle of incidence?

Exercise:

Problem:

A pencil flashlight submerged in water sends a light beam toward the surface at an angle of incidence of 30°. What is the angle of refraction in air?

Solution:

42°

Exercise:

Problem:

Light rays from the Sun make a 30° angle to the vertical when seen from below the surface of a body of water. At what angle above the horizon is the Sun?

Exercise:

Problem:

The path of a light beam in air goes from an angle of incidence of 35° to an angle of refraction of 22° when it enters a rectangular block of plastic. What is the index of refraction of the plastic?

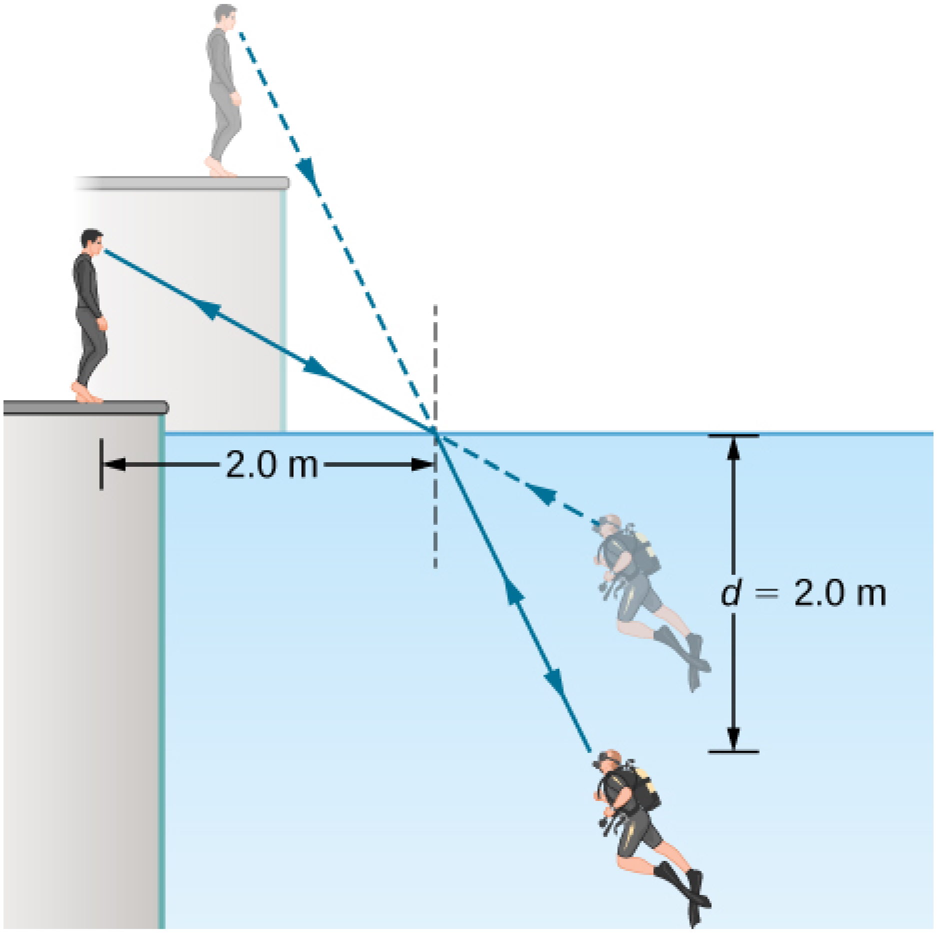
Solution:

1.53

Exercise:

Problem:

A scuba diver training in a pool looks at his instructor as shown below. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is 25.0°.



Exercise:

Problem:

(a) Using information in the preceding problem, find the height of the instructor’s head above the water, noting that you will first have to calculate the angle of incidence. (b) Find the apparent depth of the diver’s head below water as seen by the instructor.

Solution:

a. 2.9 m; b. 1.4 m

Glossary

law of refraction

when a light ray crosses from one medium to another, it changes direction by an amount that depends on the index of refraction of each medium and the sines of the angle of incidence and angle of refraction

refraction
changing of a light ray’s direction when it passes through variations in matter

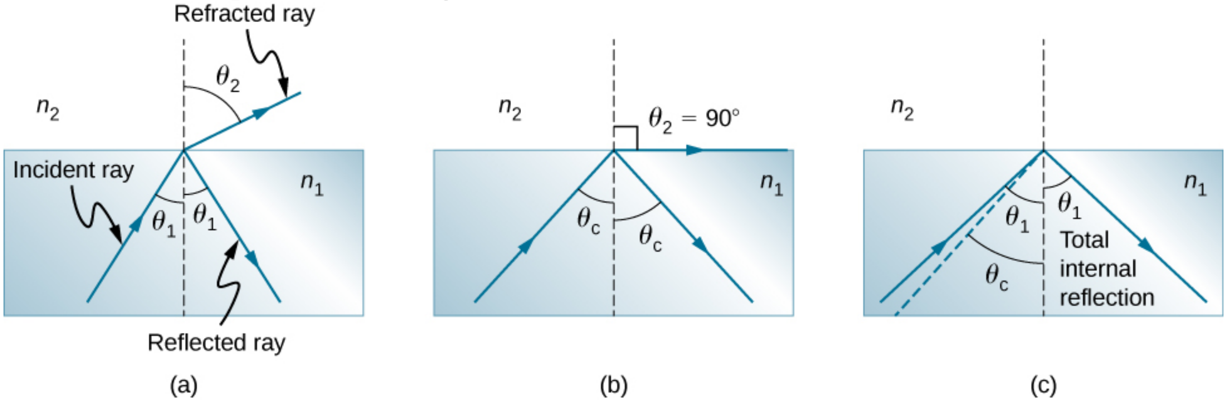
Total Internal Reflection

By the end of this section, you will be able to:

- Explain the phenomenon of total internal reflection
- Describe the workings and uses of optical fibers
- Analyze the reason for the sparkle of diamonds

A good-quality mirror may reflect more than 90% of the light that falls on it, absorbing the rest. But it would be useful to have a mirror that reflects all of the light that falls on it. Interestingly, we can produce total reflection using an aspect of refraction.

Consider what happens when a ray of light strikes the surface between two materials, as shown in [link](a). Part of the light crosses the boundary and is refracted; the rest is reflected. If, as shown in the figure, the index of refraction for the second medium is less than for the first, the ray bends away from the perpendicular. (Since $n_1 > n_2$, the angle of refraction is greater than the angle of incidence—that is, $\theta_2 > \theta_1$.) Now imagine what happens as the incident angle increases. This causes θ_2 to increase also. The largest the angle of refraction θ_2 can be is 90°, as shown in part (b). The **critical angle** θ_c for a combination of materials is defined to be the incident angle θ_1 that produces an angle of refraction of 90°. That is, θ_c is the incident angle for which $\theta_2 = 90^\circ$. If the incident angle θ_1 is greater than the critical angle, as shown in [link](c), then all of the light is reflected back into medium 1, a condition called **total internal reflection**. (As the figure shows, the reflected rays obey the law of reflection so that the angle of reflection is equal to the angle of incidence in all three cases.)



(a) A ray of light crosses a boundary where the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical angle θ_c is the angle of incidence for which the angle of refraction is 90°. (c) Total internal reflection occurs when the incident angle is greater than the critical angle.

Snell's law states the relationship between angles and indices of refraction. It is given by

Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

When the incident angle equals the critical angle $(\theta_1 = \theta_c)$, the angle of refraction is 90° $(\theta_2 = 90^\circ)$. Noting that $\sin 90^\circ = 1$, Snell's law in this case becomes

Equation:

$$n_1 \sin \theta_1 = n_2.$$

The critical angle θ_c for a given combination of materials is thus

Note:

Equation:

$$\theta_\text{c} = \sin^{-1}\left(\frac{n_2}{n_1}\right) \text{ for } n_1 > n_2.$$

Total internal reflection occurs for any incident angle greater than the critical angle θ_c, and it can only occur when the second medium has an index of refraction less than the first. Note that this equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in [link].

Example:

Determining a Critical Angle

What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air? The index of refraction for polystyrene is 1.49.

Strategy

The index of refraction of air can be taken to be 1.00, as before. Thus, the condition that the second medium (air) has an index of refraction less than the first (plastic) is satisfied, and we can use the equation

Equation:

$$\theta_{\text{c}} = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

to find the critical angle θ_{c}, where $n_2 = 1.00$ and $n_1 = 1.49$.

Solution

Substituting the identified values gives

Equation:

$$\theta_{\text{c}} = \sin^{-1}\left(\frac{1.00}{1.49}\right) = \sin^{-1}(0.671) = 42.2^\circ.$$

Significance

This result means that any ray of light inside the plastic that strikes the surface at an angle greater than 42.2° is totally reflected. This makes the inside surface of the clear plastic a perfect mirror for such rays, without any need for the silvering used on common mirrors. Different combinations of materials have different critical angles, but any combination with $n_1 > n_2$ can produce total internal reflection. The same calculation as made here shows that the critical angle for a ray going from water to air is 48.6°, whereas that from diamond to air is 24.4°, and that from flint glass to crown glass is 66.3°.

Note:
Exercise:

Problem:

Check Your Understanding At the surface between air and water, light rays can go from air to water and from water to air. For which ray is there no possibility of total internal reflection?

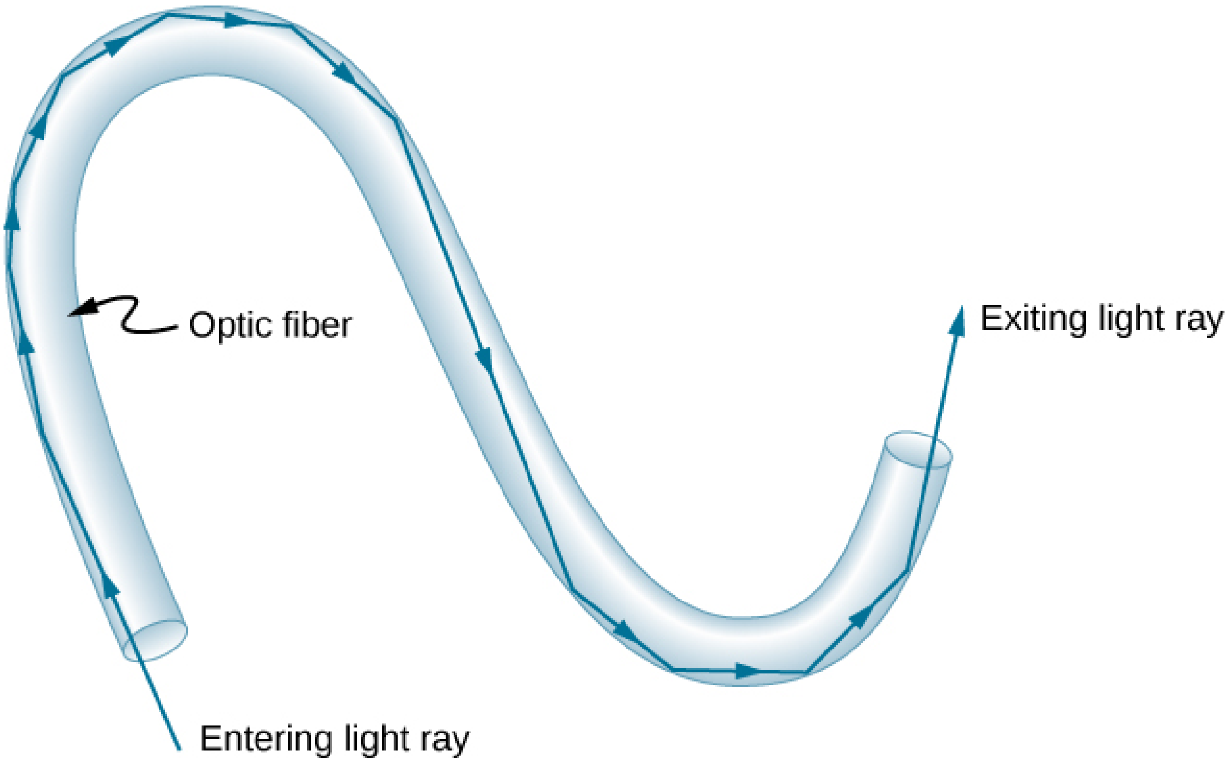
Solution:

air to water, because the condition that the second medium must have a smaller index of refraction is not satisfied

In the photo that opens this chapter, the image of a swimmer underwater is captured by a camera that is also underwater. The swimmer in the upper half of the photograph, apparently facing upward, is, in fact, a reflected image of the swimmer below. The circular ripple near the photograph's center is actually on the water surface. The undisturbed water surrounding it makes a good reflecting surface when viewed from below, thanks to total internal reflection. However, at the very top edge of this photograph, rays from below strike the surface with incident angles less than the critical angle, allowing the camera to capture a view of activities on the pool deck above water.

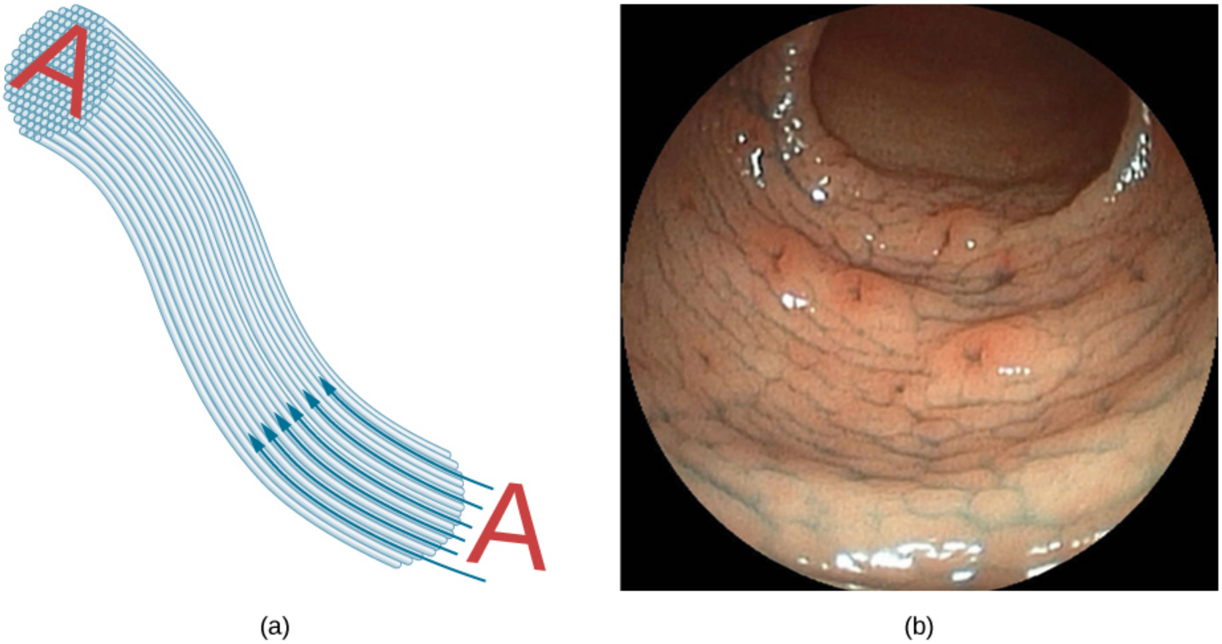
Fiber Optics: Endoscopes to Telephones

Fiber optics is one application of total internal reflection that is in wide use. In communications, it is used to transmit telephone, internet, and cable TV signals. **Fiber optics** employs the transmission of light down fibers of plastic or glass. Because the fibers are thin, light entering one is likely to strike the inside surface at an angle greater than the critical angle and, thus, be totally reflected ([link]). The index of refraction outside the fiber must be smaller than inside. In fact, most fibers have a varying refractive index to allow more light to be guided along the fiber through total internal refraction. Rays are reflected around corners as shown, making the fibers into tiny light pipes.



Light entering a thin optic fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection and incidence remain large.

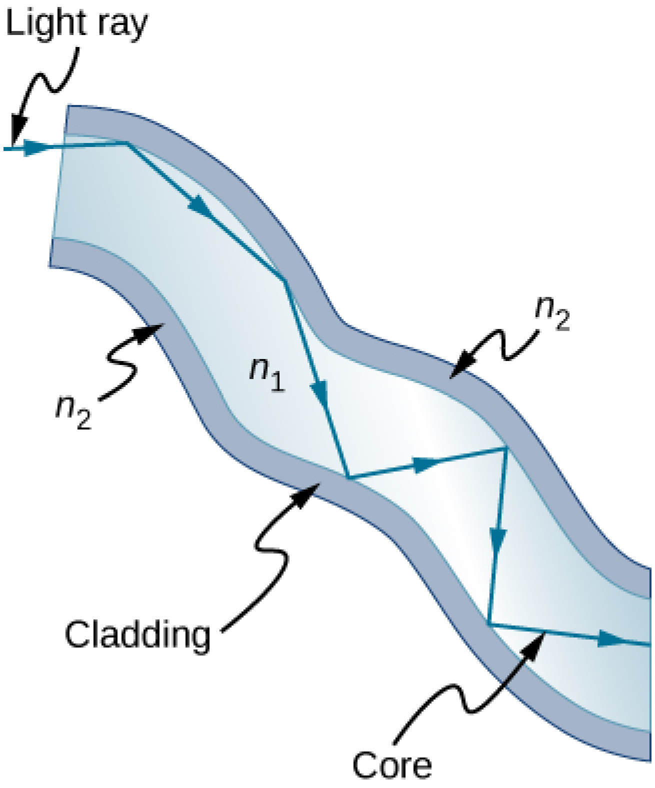
Bundles of fibers can be used to transmit an image without a lens, as illustrated in [link]. The output of a device called an endoscope is shown in [link](b). Endoscopes are used to explore the interior of the body through its natural orifices or minor incisions. Light is transmitted down one fiber bundle to illuminate internal parts, and the reflected light is transmitted back out through another bundle to be observed.



(a) An image “A” is transmitted by a bundle of optical fibers. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown of a human epiglottis (a structure at the base of the tongue). (credit b: modification of work by “Med_Chaos”/Wikimedia Commons)

Fiber optics has revolutionized surgical techniques and observations within the body, with a host of medical diagnostic and therapeutic uses. Surgery can be performed, such as arthroscopic surgery on a knee or shoulder joint, employing cutting tools attached to and observed with the endoscope. Samples can also be obtained, such as by lassoing an intestinal polyp for external examination. The flexibility of the fiber optic bundle allows doctors to navigate it around small and difficult-to-reach regions in the body, such as the intestines, the heart, blood vessels, and joints. Transmission of an intense laser beam to burn away obstructing plaques in major arteries, as well as delivering light to activate chemotherapy drugs, are becoming commonplace. Optical fibers have in fact enabled microsurgery and remote surgery where the incisions are small and the surgeon’s fingers do not need to touch the diseased tissue.

Optical fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core ([link]). The cladding prevents light from being transmitted between fibers in a bundle. Without cladding, light could pass between fibers in contact, since their indices of refraction are identical. Since no light gets into the cladding (there is total internal reflection back into the core), none can be transmitted between clad fibers that are in contact with one another. Instead, the light is propagated along the length of the fiber, minimizing the loss of signal and ensuring that a quality image is formed at the other end. The cladding and an additional protective layer make optical fibers durable as well as flexible.



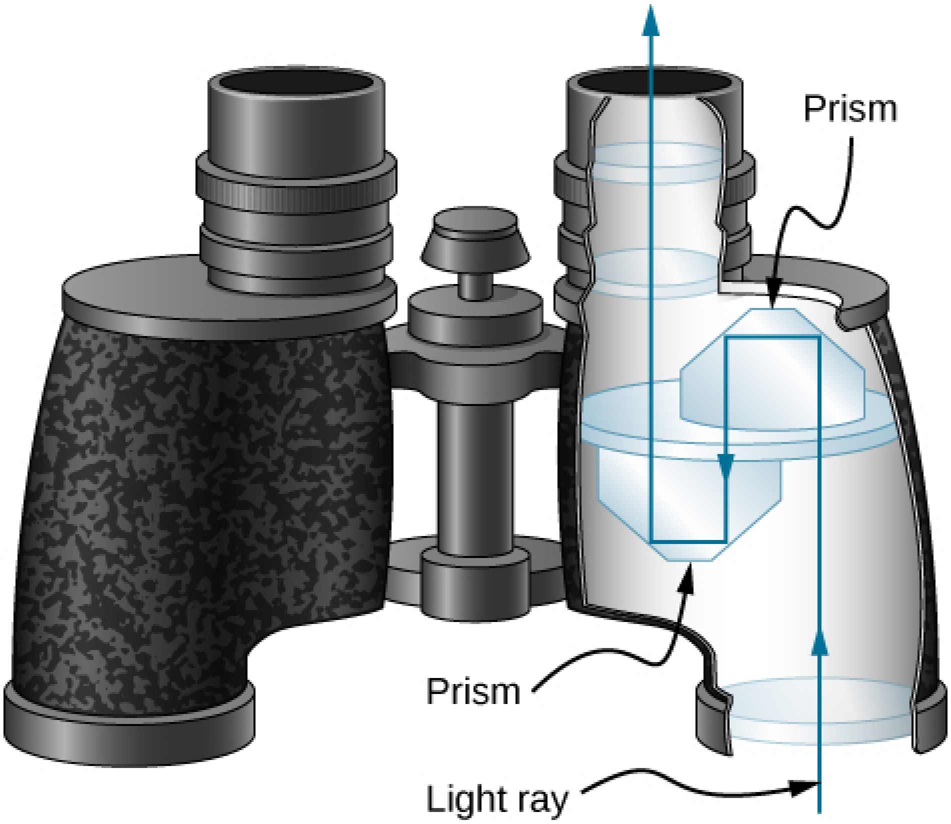
Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another.

Special tiny lenses that can be attached to the ends of bundles of fibers have been designed and fabricated. Light emerging from a fiber bundle can be focused through such a lens, imaging a tiny spot. In some cases, the spot can be scanned, allowing quality imaging of a region inside the body. Special minute optical filters inserted at the end of the fiber bundle have the capacity to image the interior of organs located tens of microns below the surface without cutting the surface—an area known as nonintrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

In another type of application, optical fibers are commonly used to carry signals for telephone conversations and internet communications. Extensive optical fiber cables have been placed on the ocean floor and underground to enable optical communications. Optical fiber communication systems offer several advantages over electrical (copper)-based systems, particularly for long distances. The fibers can be made so transparent that light can travel many kilometers before it becomes dim enough to require amplification—much superior to copper conductors. This property of optical fibers is called low loss. Lasers emit light with characteristics that allow far more conversations in one fiber than are possible with electric signals on a single conductor. This property of optical fibers is called high bandwidth. Optical signals in one fiber do not produce undesirable effects in other adjacent fibers. This property of optical fibers is called reduced crosstalk. We shall explore the unique characteristics of laser radiation in a later chapter.

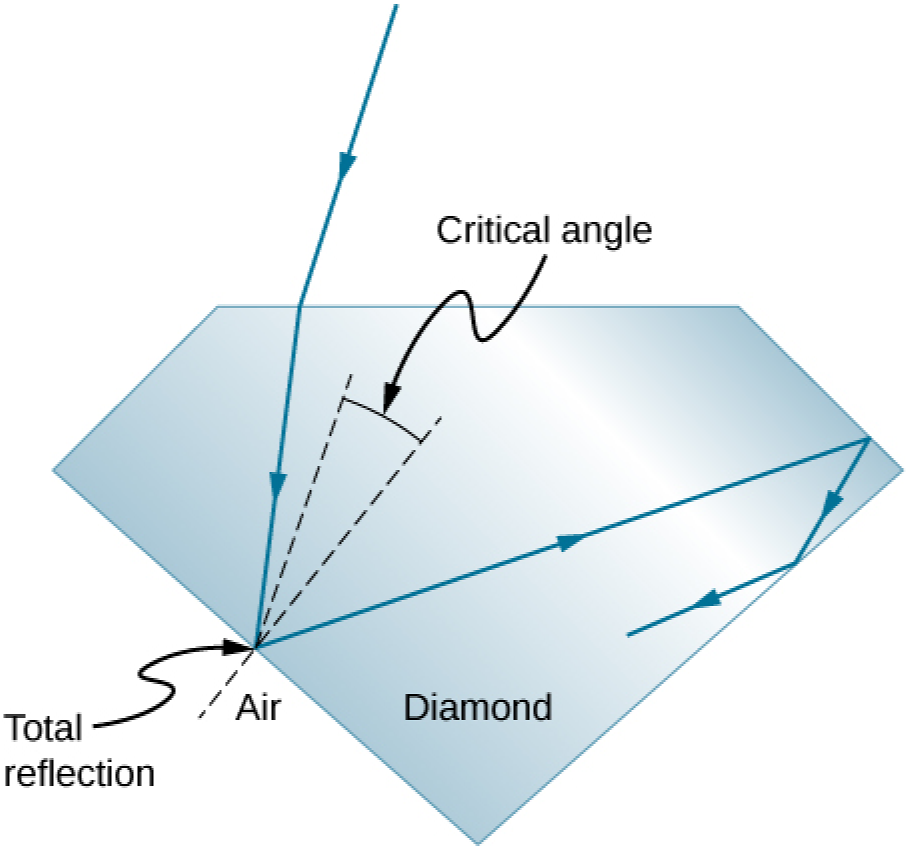
Corner Reflectors and Diamonds

Corner reflectors (The Law of Reflection) are perfectly efficient when the conditions for total internal reflection are satisfied. With common materials, it is easy to obtain a critical angle that is less than 45°. One use of these perfect mirrors is in binoculars, as shown in [link]. Another use is in periscopes found in submarines.



These binoculars employ corner reflectors (prisms) with total internal reflection to get light to the observer’s eyes.

Total internal reflection, coupled with a large index of refraction, explains why diamonds sparkle more than other materials. The critical angle for a diamond-to-air surface is only 24.4°, so when light enters a diamond, it has trouble getting back out ([link]). Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4°. Facets on diamonds are specifically intended to make this unlikely. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated before exiting—hence the bright sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but it is not as large as diamond, so it is not as highly prized. Cubic zirconia is manufactured and has an even higher index of refraction (≈ 2.17), but it is still less than that of diamond.) The colors you see emerging from a clear diamond are not due to the diamond’s color, which is usually nearly colorless. The colors result from dispersion, which we discuss in Dispersion. Colored diamonds get their color from structural defects of the crystal lattice and the inclusion of minute quantities of graphite and other materials. The Argyle Mine in Western Australia produces around 90% of the world’s pink, red, champagne, and cognac diamonds, whereas around 50% of the world’s clear diamonds come from central and southern Africa.



Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle brightly.

Note:
Explore refraction and reflection of light between two media with different indices of refraction. Try to make the refracted ray disappear with total internal reflection. Use the protractor tool to measure the critical angle and compare with the prediction from [link].

Summary

- The incident angle that produces an angle of refraction of 90° is called the critical angle.
- Total internal reflection is a phenomenon that occurs at the boundary between two media, such that if the incident angle in the first medium is greater than the critical angle, then all the light is reflected back into that medium.
- Fiber optics involves the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection.
- Cladding prevents light from being transmitted between fibers in a bundle.
- Diamonds sparkle due to total internal reflection coupled with a large index of refraction.

Conceptual Questions

Exercise:

Problem:

A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.

Solution:

The gemstone becomes invisible when its index of refraction is the same, or at least similar to, the water surrounding it. Because diamond has a particularly high index of refraction, it can still sparkle as a result of total internal reflection, not invisible.

Exercise:

Problem:

The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.

Exercise:

Problem:

How can you use total internal reflection to estimate the index of refraction of a medium?

Solution:

One can measure the critical angle by looking for the onset of total internal reflection as the angle of incidence is varied. [link] can then be applied to compute the index of refraction.

Problems

Exercise:

Problem:

Verify that the critical angle for light going from water to air is 48.6°, as discussed at the end of [link], regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

Exercise:

Problem:

(a) At the end of [link], it was stated that the critical angle for light going from diamond to air is 24.4°. Verify this. (b) What is the critical angle for light going from zircon to air?

Solution:

a. 24.42°; b. 31.33°

Exercise:

Problem:

An optical fiber uses flint glass clad with crown glass. What is the critical angle?

Exercise:

Problem:

At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?

Solution:

79.11°

Exercise:

Problem:

Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0°, what must be the minimum index of refraction of the material from which the reflector is made?

Exercise:

Problem:

You can determine the index of refraction of a substance by determining its critical angle. (a) What is the index of refraction of a substance that has a critical angle of 68.4° when submerged in water? What is the substance, based on [link]? (b) What would the critical angle be for this substance in air?

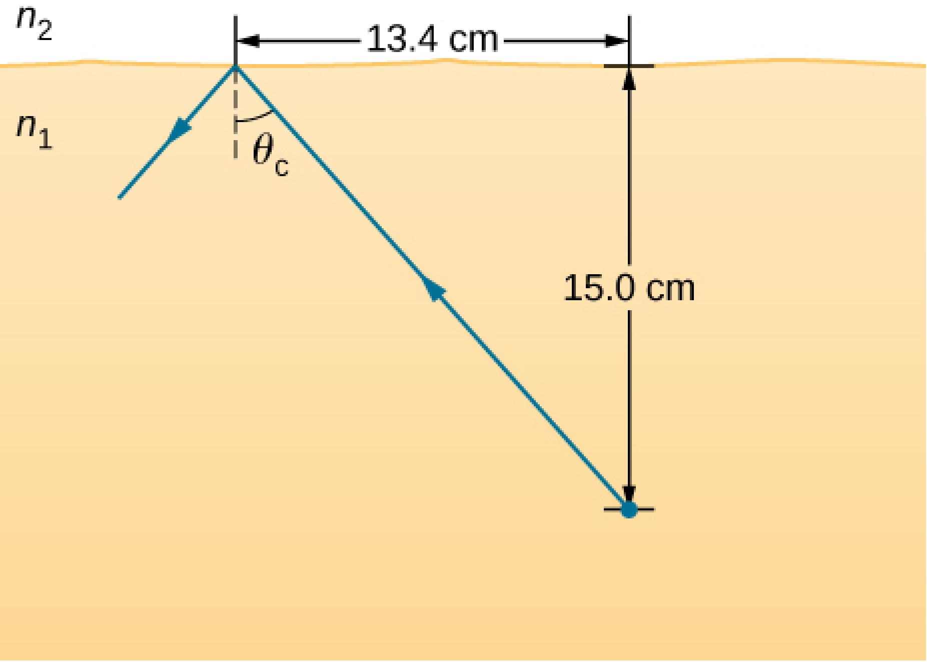
Solution:

a. 1.43, fluorite; b. 44.2°

Exercise:

Problem:

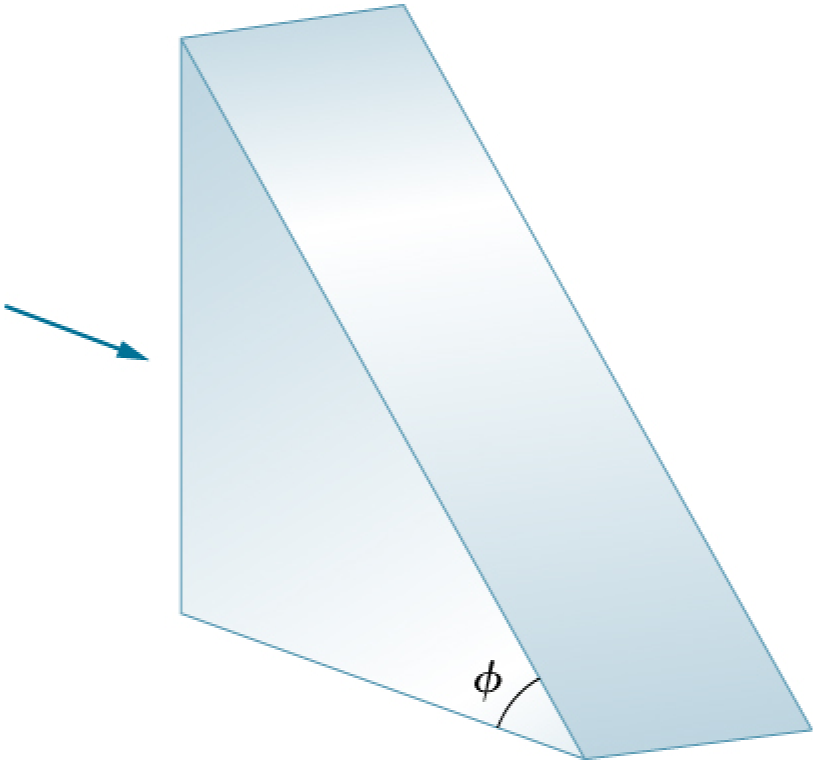
A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown below. What is the index of refraction for the liquid and its likely identification?



Exercise:

Problem:

Light rays fall normally on the vertical surface of the glass prism $(n = 1.50)$ shown below. (a) What is the largest value for ϕ such that the ray is totally reflected at the slanted face? (b) Repeat the calculation of part (a) if the prism is immersed in water.



Solution:

a. 48.2°; b. 27.3°

Glossary

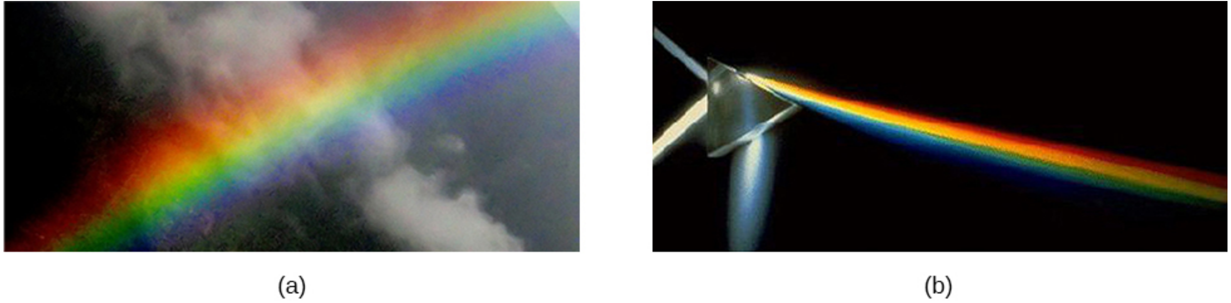
critical angle	incident angle that produces an angle of refraction of 90°
fiber optics	field of study of the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection
total internal reflection	phenomenon at the boundary between two media such that all the light is reflected and no refraction occurs

Dispersion

By the end of this section, you will be able to:

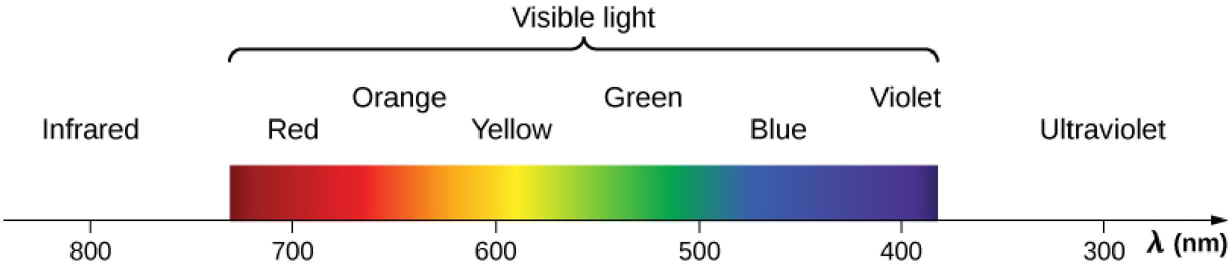
- Explain the cause of dispersion in a prism
- Describe the effects of dispersion in producing rainbows
- Summarize the advantages and disadvantages of dispersion

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond ([link]).


(a)
(b)

The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit a: modification of work by “Alfredo55”/Wikimedia Commons; credit b: modification of work by NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. These colors are associated with different wavelengths of light, as shown in [link]. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye’s response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow, because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors shown in the figure. This implies that white light is spread out in a rainbow according to wavelength. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever the propagation of light depends on wavelength.



Even though rainbows are associated with six colors, the rainbow is a continuous distribution of colors according to wavelengths.

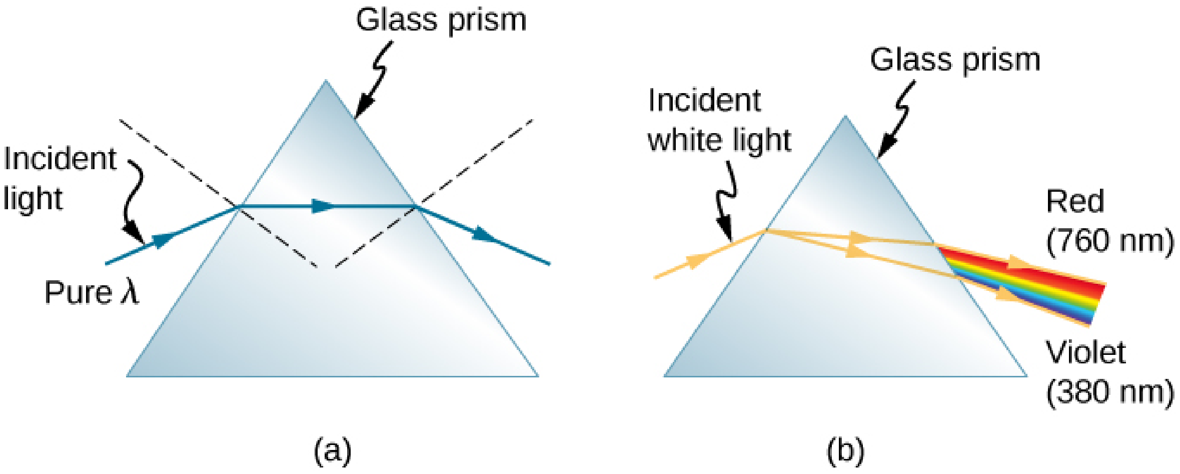
Any type of wave can exhibit dispersion. For example, sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called interstellar medium.

Note:
Nick Moore's video discusses dispersion of a pulse as he taps a long spring. Follow his explanation as Moore replays the high-speed footage showing high frequency waves outrunning the lower frequency waves.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we know from Snell's law. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength ([link]). Note that for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus, violet light is bent more than red light, as shown for a prism in [link](b). White light is dispersed into the same sequence of wavelengths as seen in [link] and [link].

Medium	**Red (660 nm)**	**Orange (610 nm)**	**Yellow (580 nm)**	**Green (550 nm)**	**Blue (470 nm)**	**Violet (410 nm)**
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

Index of Refraction *n* in Selected Media at Various Wavelengths

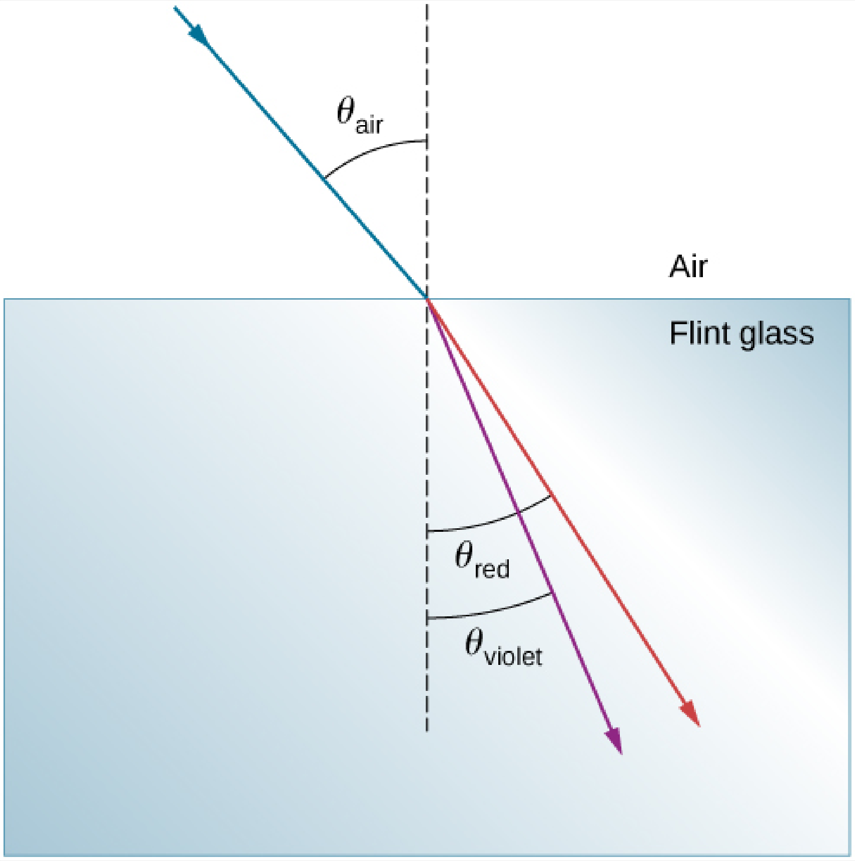

Glass prism
Incident light
Pure λ
(a)
Incident white light
Glass prism
Red (760 nm)
Violet (380 nm)
(b)

(a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Example:

Dispersion of White Light by Flint Glass

A beam of white light goes from air into flint glass at an incidence angle of 43.2°. What is the angle between the red (660 nm) and violet (410 nm) parts of the refracted light?



Strategy

Values for the indices of refraction for flint glass at various wavelengths are listed in [link]. Use these values for calculate the angle of refraction for each color and then take the difference to find the dispersion angle.

Solution

Applying the law of refraction for the red part of the beam

Equation:

$$n_{\text{air}} \sin \theta_{\text{air}} = n_{\text{red}} \sin \theta_{\text{red}},$$

we can solve for the angle of refraction as

Equation:

$$\theta_{\text{red}} = \sin^{-1}\left(\frac{n_{\text{air}} \sin \theta_{\text{air}}}{n_{\text{red}}}\right) = \sin^{-1}\left[\frac{(1.000) \sin 43.2^\circ}{(1.662)}\right] = 27.0^\circ.$$

Similarly, the angle of incidence for the violet part of the beam is
Equation:

$$\theta_{\text{violet}} = \sin^{-1}\left(\frac{n_{\text{air}} \sin \theta_{\text{air}}}{n_{\text{violet}}}\right) = \sin^{-1}\left[\frac{(1.000) \sin 43.2^\circ}{(1.698)}\right] = 26.4^\circ.$$

The difference between these two angles is
Equation:

$$\theta_{\text{red}} - \theta_{\text{violet}} = 27.0^\circ - 26.4^\circ = 0.6^\circ.$$

Significance
Although 0.6° may seem like a negligibly small angle, if this beam is allowed to propagate a long enough distance, the dispersion of colors becomes quite noticeable.

Note:
Exercise:

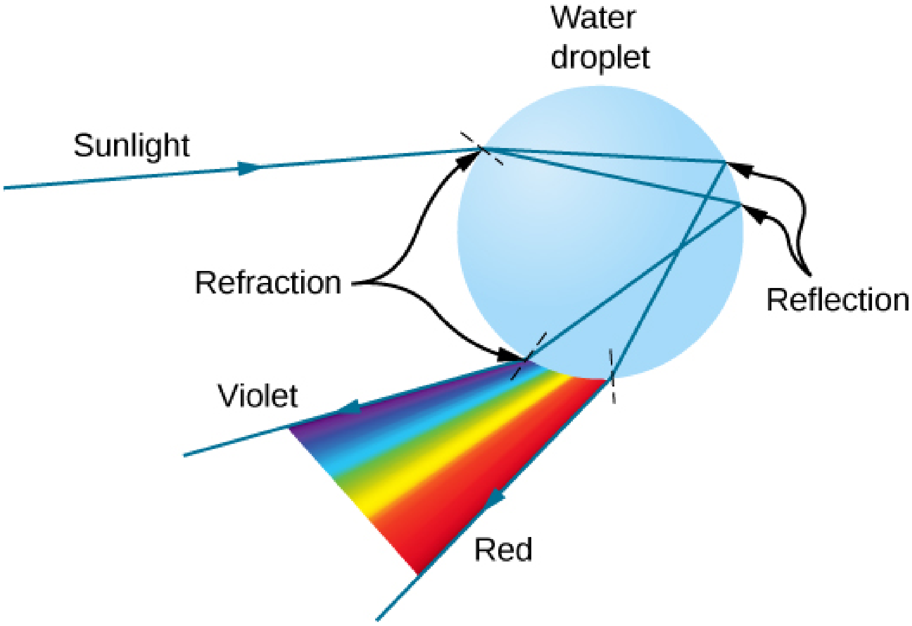
Problem:

Check Your Understanding In the preceding example, how much distance inside the block of flint glass would the red and the violet rays have to progress before they are separated by 1.0 mm?

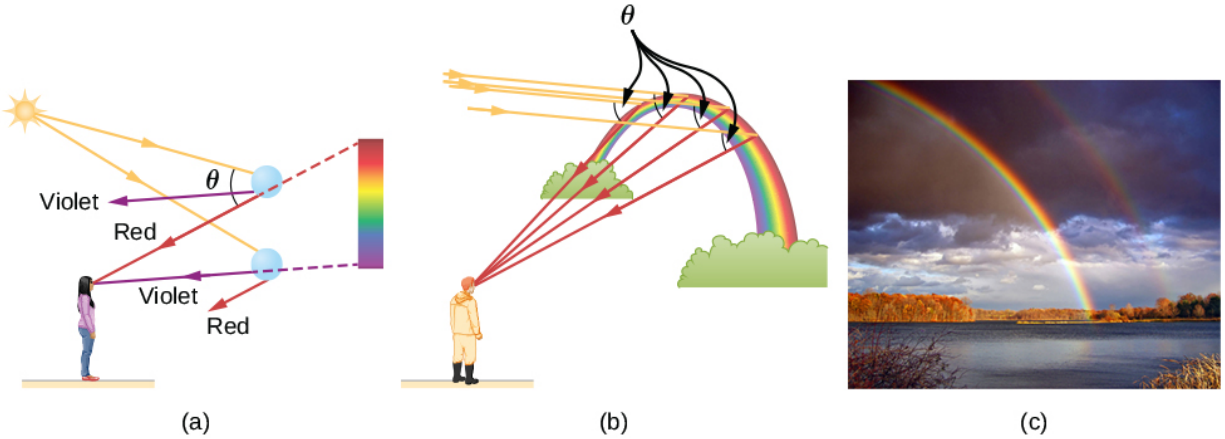
Solution:

9.3 cm

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the Sun. Light enters a drop of water and is reflected from the back of the drop ([link]). The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed ([link](a)). (No dispersion occurs at the back surface, because the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the Sun, as illustrated in part (b). If two reflections of light occur within the water drop, another “secondary” rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc, as in part (c), and produces colors in the reverse order of the primary rainbow, with red at the lowest angle and violet at the largest angle.


Water droplet
Sunlight
Refraction
Reflection
Violet
Red

A ray of light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.


θ
Violet
Red
Violet
Red
θ
(a)
(b)
(c)

(a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight for the observer to receive the refracted rays. (c) Double rainbow. (credit c: modification of work by "Nicholas"/Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through.

Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection, and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Conceptual Questions

Exercise:

Problem:

Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to that shown below. Some of us have seen the formation of a double rainbow; is it physically possible to observe a triple rainbow?

(credit: "Chad"/Flickr)

Exercise:

Problem:

A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.

Solution:

In addition to total internal reflection, rays that refract into and out of diamond crystals are subject to dispersion due to varying values of n across the spectrum, resulting in a sparkling display of colors.

Problems

Exercise:

Problem:

(a) What is the ratio of the speed of red light to violet light in diamond, based on [link]? (b) What is this ratio in polystyrene? (c) Which is more dispersive?

Exercise:

Problem:

A beam of white light goes from air into water at an incident angle of 75.0°. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?

Solution:

46.5° for red, 46.0° for violet

Exercise:

Problem:

By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?

Exercise:

Problem:

(a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a 30.0° incident angle. What is the angle between the colors when they emerge? (b) How far would they have to travel to be separated by 1.00 mm?

Solution:

a. 0.04°; b. 1.3 m

Exercise:

Problem:

A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a 60.0° incident angle. What is the angle between the two colors in water?

Exercise:

Problem:

A ray of 610-nm light goes from air into fused quartz at an incident angle of 55.0°. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?

Solution:

72.8°

Exercise:

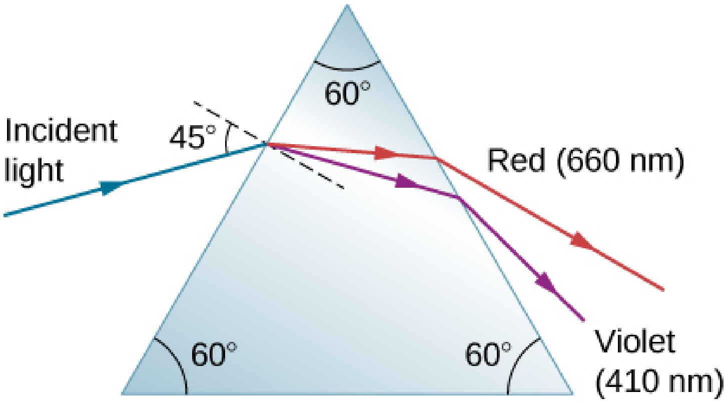
Problem:

A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00-cm-thick flat piece of crown glass and back to air again. The beam strikes at a 30.0° incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?

Exercise:

Problem:

A narrow beam of white light enters a prism made of crown glass at a 45.0° incident angle, as shown below. At what angles, θ_R and θ_V, do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?



Solution:

53.5° for red, 55.2° for violet

Glossary

dispersion
spreading of light into its spectrum of wavelengths

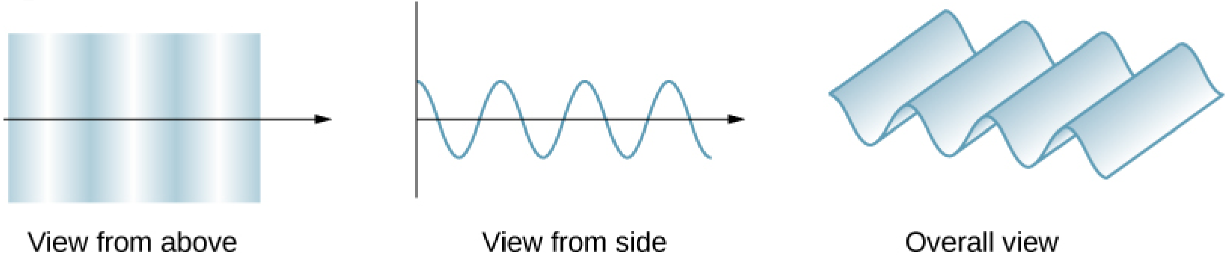
Huygens's Principle

By the end of this section, you will be able to:

- Describe Huygens's principle
- Use Huygens's principle to explain the law of reflection
- Use Huygens's principle to explain the law of refraction
- Use Huygens's principle to explain diffraction

So far in this chapter, we have been discussing optical phenomena using the ray model of light. However, some phenomena require analysis and explanations based on the wave characteristics of light. This is particularly true when the wavelength is not negligible compared to the dimensions of an optical device, such as a slit in the case of *diffraction*. Huygens's principle is an indispensable tool for this analysis.

[link] shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wave fronts (or wave crests) as if we were looking down on ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps more useful in developing concepts about **wave optics**.



View from above

View from side

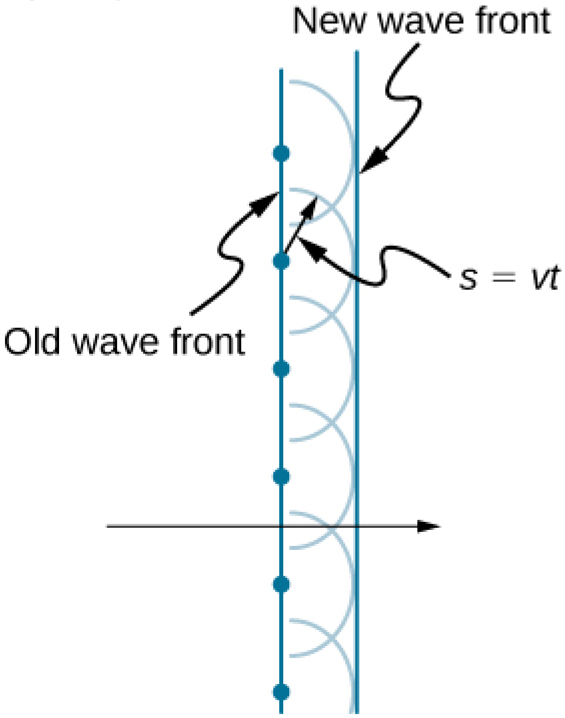
Overall view

A transverse wave, such as an electromagnetic light wave, as viewed from above and from the side. The direction of propagation is perpendicular to the wave fronts (or wave crests) and is represented by a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate.

Starting from some known position, **Huygens's principle** states that every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.

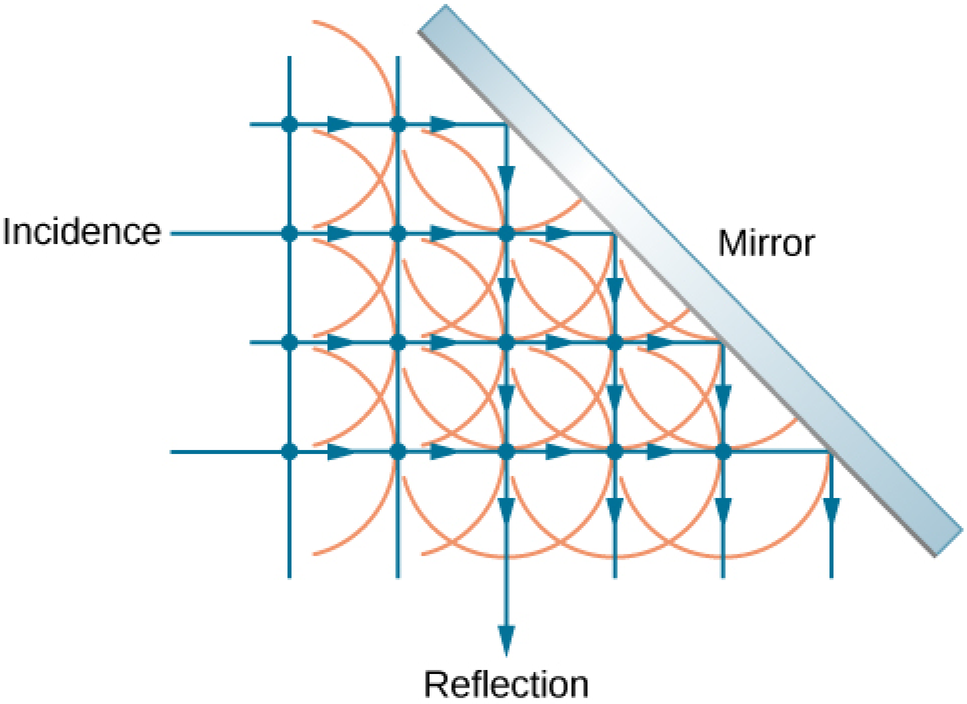
[link] shows how Huygens's principle is applied. A wave front is the long edge that moves, for example, with the crest or the trough. Each point on the wave front emits a semicircular wave that moves at the propagation speed *v*. We can draw these wavelets at a time *t* later, so that they have moved a distance $s = vt$. The new wave front is a plane tangent to the wavelets and is where we would expect the wave to be a time *t* later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. It is useful not only in describing how light waves propagate but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.



Huygens's principle applied to a straight wave front. Each point on the wave front emits a semicircular wavelet that moves a distance $s = vt$. The new wave front is a line tangent to the wavelets.

Reflection

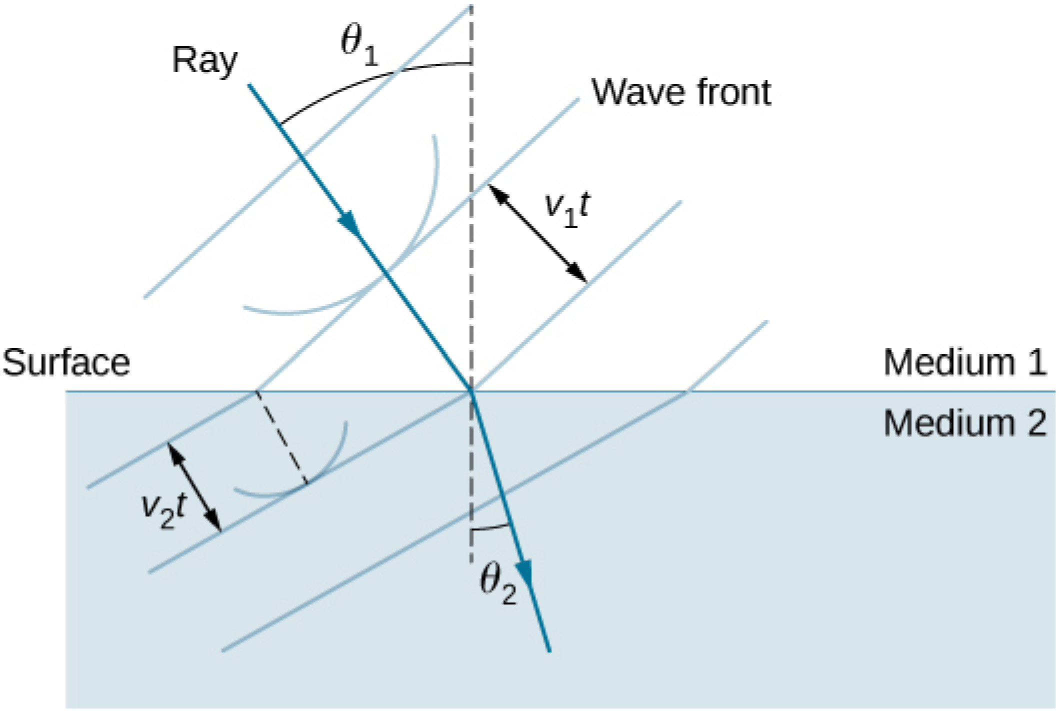
[link] shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wave front strikes the mirror, wavelets are first emitted from the left part of the mirror and then from the right. The wavelets closer to the left have had time to travel farther, producing a wave front traveling in the direction shown.



Huygens’s principle applied to a plane wave front striking a mirror. The wavelets shown were emitted as each point on the wave front struck the mirror. The tangent to these wavelets shows that the new wave front has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wave front, as shown by the downward-pointing arrows.

Refraction

The law of refraction can be explained by applying Huygens’s principle to a wave front passing from one medium to another ([link]). Each wavelet in the figure was emitted when the wave front crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wave front changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell’s law can be derived from the geometry in [link] ([link]).



Huygens’s principle applied to a plane wave front traveling from one medium to another, where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

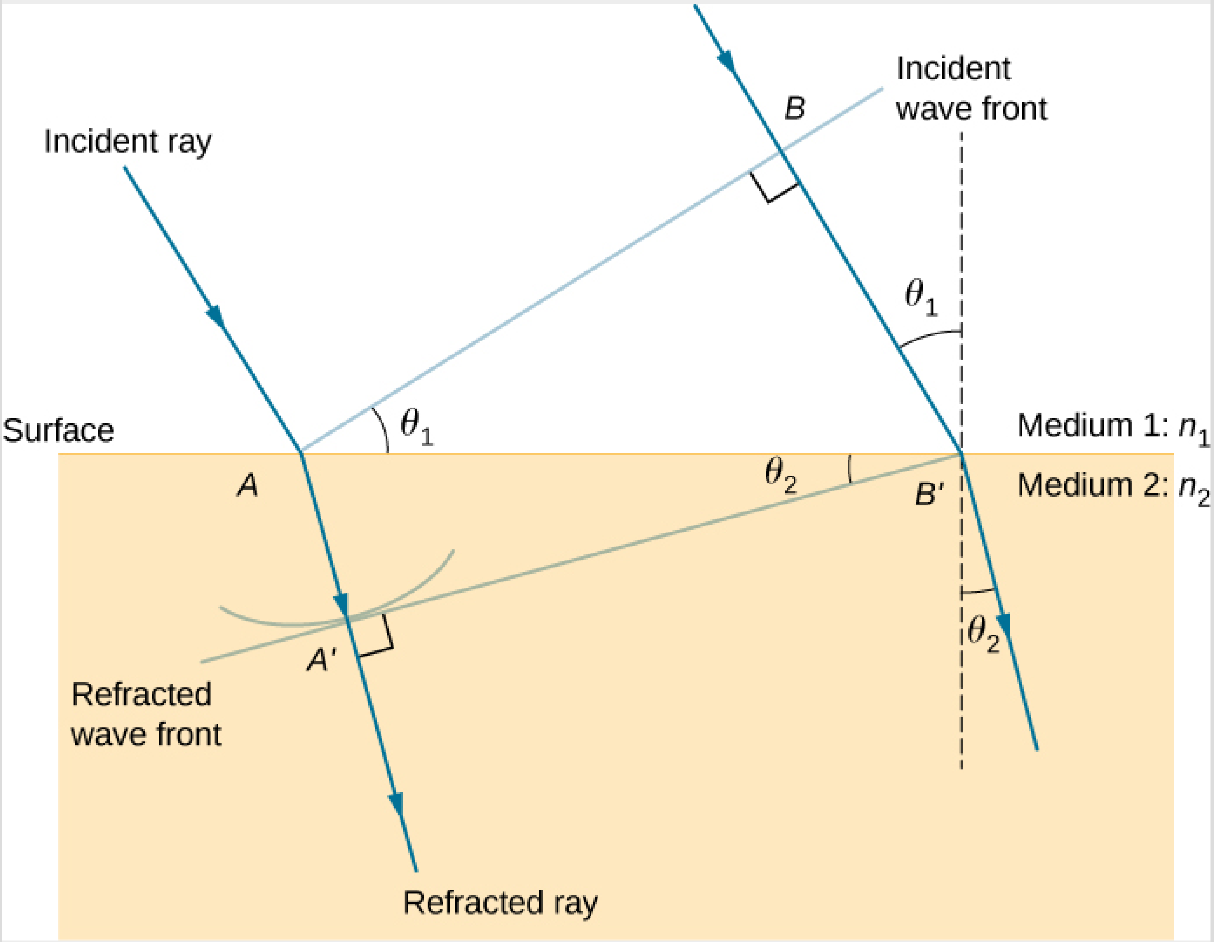
Example:

Deriving the Law of Refraction

By examining the geometry of the wave fronts, derive the law of refraction.

Strategy

Consider [link], which expands upon [link]. It shows the incident wave front just reaching the surface at point A, while point B is still well within medium 1. In the time Δt it takes for a wavelet from B to reach $B\prime$ on the surface at speed $v_1 = c/n_1$, a wavelet from A travels into medium 2 a distance of $AA\prime = v_2 \Delta t$, where $v_2 = c/n_2$. Note that in this example, v_2 is slower than v_1 because $n_1 < n_2$.


Incident wave front
B
Incident ray
θ_1
θ_1
Surface
Medium 1: n_1
θ_2
A
B′
Medium 2: n_2
θ_2
A′
Refracted wave front
Refracted ray

Geometry of the law of refraction from medium 1 to medium 2.

Solution

The segment on the surface AB' is shared by both the triangle ABB' inside medium 1 and the triangle $AA'B'$ inside medium 2. Note that from the geometry, the angle $\angle BAB'$ is equal to the angle of incidence, θ_1. Similarly, $\angle AB'A'$ is θ_2.

The length of AB' is given in two ways as

Equation:

$$AB' = \frac{BB'}{\sin\theta_1} = \frac{AA'}{\sin\theta_2}.$$

Inverting the equation and substituting $AA' = c\Delta t/n_2$ from above and similarly $BB' = c\Delta t/n_1$, we obtain

Equation:

$$\frac{\sin\theta_1}{c\Delta t/n_1} = \frac{\sin\theta_2}{c\Delta t/n_2}.$$

Cancellation of $c\Delta t$ allows us to simplify this equation into the familiar form

Equation:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

Significance

Although the law of refraction was established experimentally by Snell and stated in Refraction, its derivation here requires Huygens's principle and the understanding that the speed of light is different in different media.

Note:

Exercise:

Problem:

Check Your Understanding In [link], we had $n_1 < n_2$. If n_2 were decreased such that $n_1 > n_2$ and the speed of light in medium 2 is faster than in medium 1, what would happen to the length of $AA\prime$? What would happen to the wave front $A\prime B\prime$ and the direction of the refracted ray?

Solution:

$AA\prime$ becomes longer, $A\prime B\prime$ tilts further away from the surface, and the refracted ray tilts away from the normal.

Note:
This applet by Walter Fendt shows an animation of reflection and refraction using Huygens's wavelets while you control the parameters. Be sure to click on "Next step" to display the wavelets. You can see the reflected and refracted wave fronts forming.

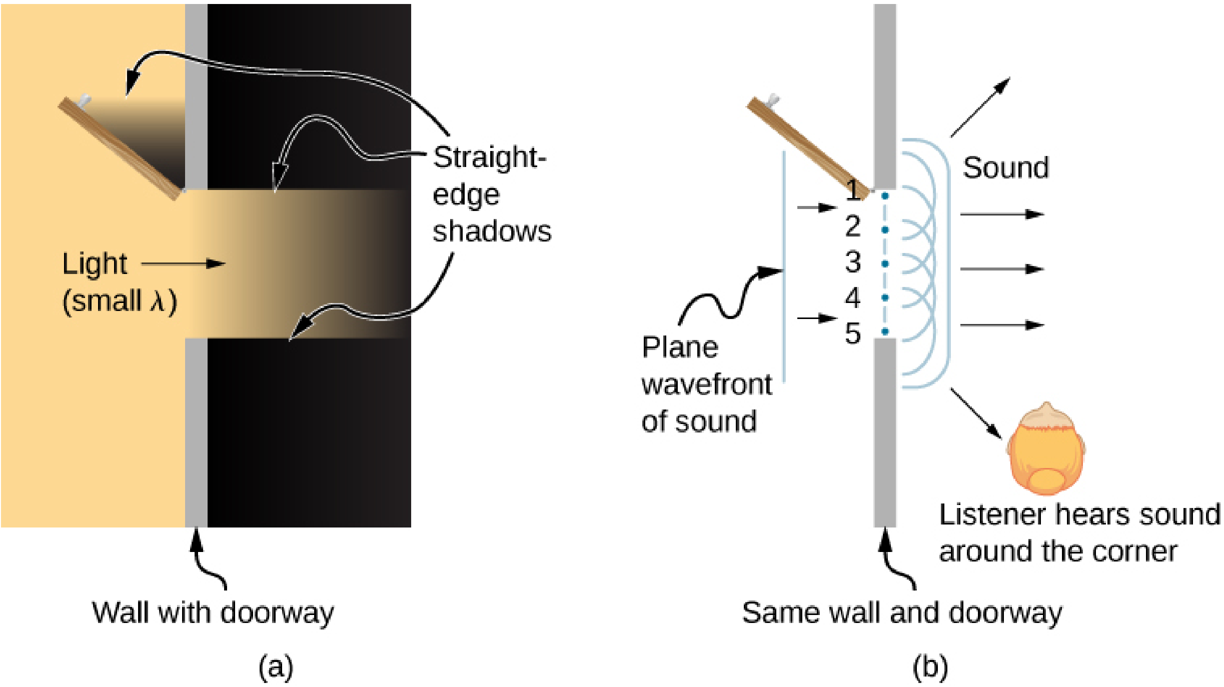
Diffraction

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we observe a sharp shadow of the doorway on the floor of the room, and no visible light bends around corners into other parts of the room. When sound passes through a door, we hear it everywhere in the room and thus observe that sound spreads out when passing through such an opening ([link]). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz,

Equation:

$$\lambda = \frac{c}{f} = \frac{330\,\text{m/s}}{1000\,\text{s}^{-1}} = 0.33\,\text{m},$$

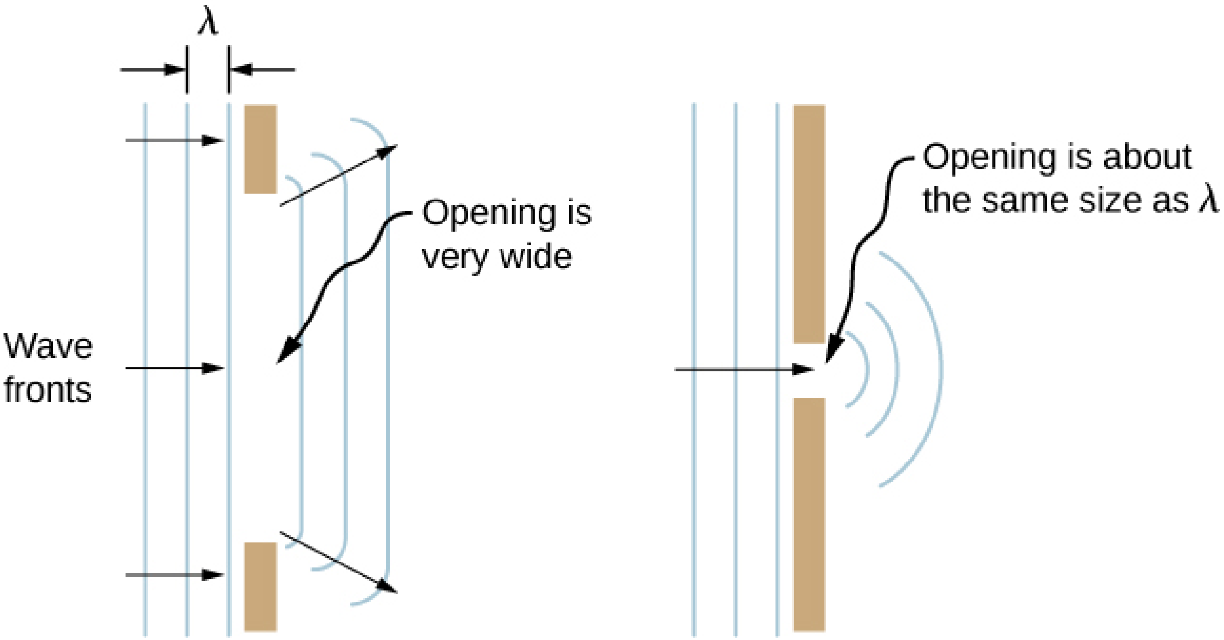
about three times smaller than the width of the doorway).



Straight-edge shadows
Light (small λ)
Wall with doorway
(a)
Sound
1
2
3
4
5
Plane wavefront of sound
Listener hears sound around the corner
Same wall and doorway
(b)

(a) Light passing through a doorway makes a sharp outline on the floor. Since light’s wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings such as slits, we can use Huygens’s principle to see that light bends as sound does ([link]). The bending of a wave around the edges of an opening or an obstacle is called diffraction. Diffraction is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus, the horizontal diffraction of the laser beam after it passes through the slits in [link] is evidence that light is a wave. You will learn about diffraction in much more detail in the chapter on Diffraction.



Huygens’s principle applied to a plane wave front striking an opening. The edges of the wave front bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

Summary

- According to Huygens’s principle, every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.

- A mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection.
- The law of refraction can be explained by applying Huygens’s principle to a wave front passing from one medium to another.
- The bending of a wave around the edges of an opening or an obstacle is called diffraction.

Conceptual Questions

Exercise:

Problem:

How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?

Exercise:

Problem: Does Huygens’s principle apply to all types of waves?

Solution:

yes

Exercise:

Problem:

If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Does the reverse hold true? That is, if diffraction is not observed, does that mean the phenomenon is not a wave?

Glossary

Huygens’s principle

every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself; the new wave front is a plane tangent to all of the wavelets

wave optics

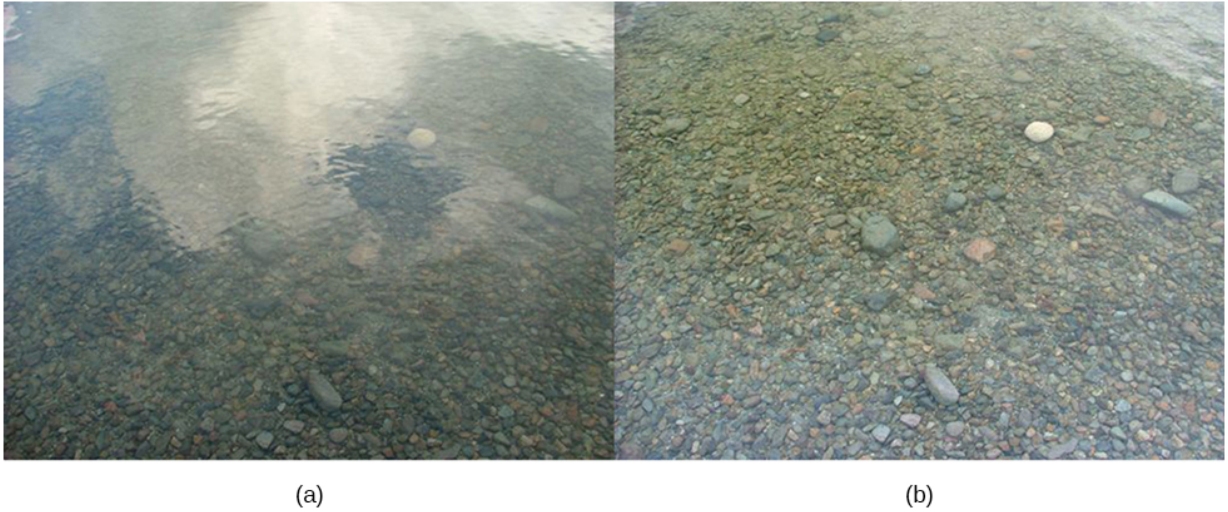
part of optics dealing with the wave aspect of light

Polarization

By the end of this section, you will be able to:

- Explain the change in intensity as polarized light passes through a polarizing filter
- Calculate the effect of polarization by reflection and Brewster’s angle
- Describe the effect of polarization by scattering
- Explain the use of polarizing materials in devices such as LCDs

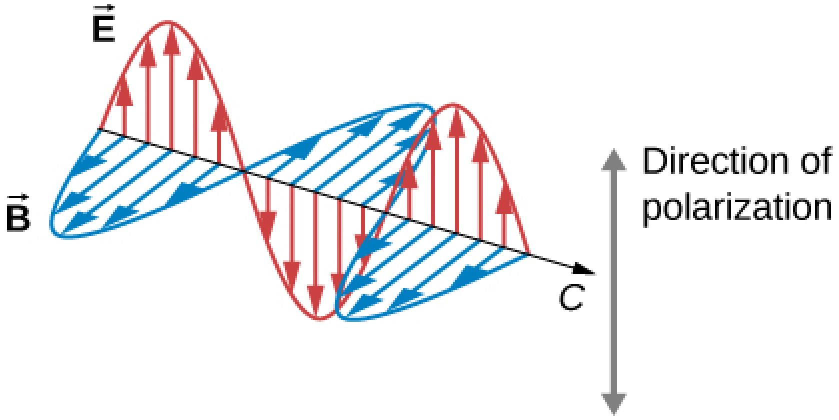
Polarizing sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass ([link]). They have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.


(a) (b)

These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit a and credit b: modifications of work by “Amithshs”/Wikimedia Commons)

Malus’s Law

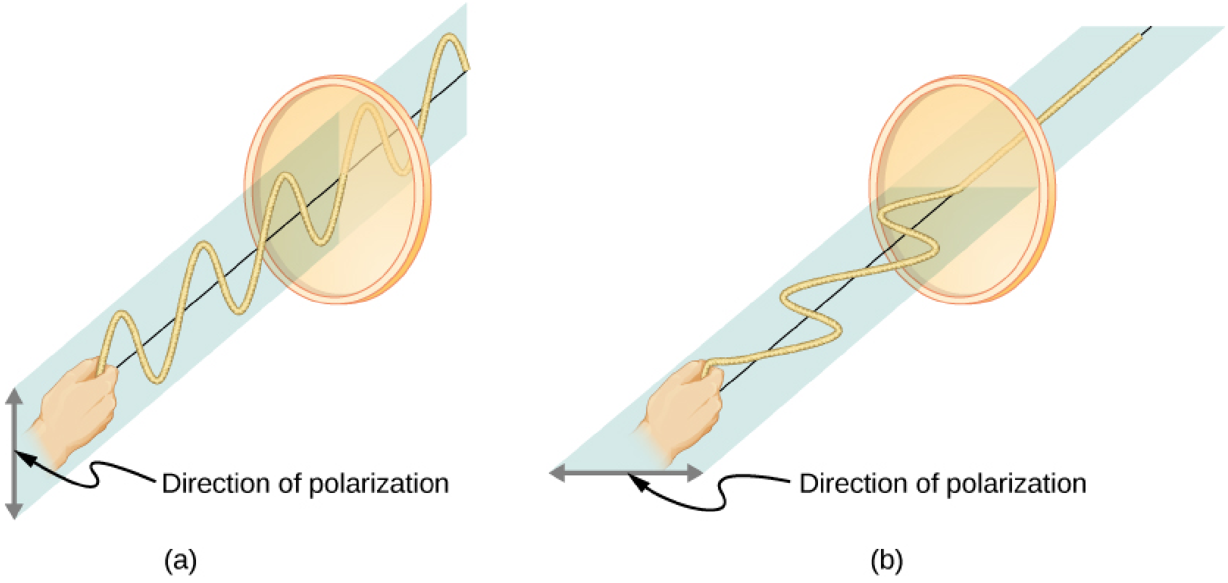
Light is one type of electromagnetic (EM) wave. As noted in the previous chapter on Electromagnetic Waves, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation ([link]). However, in general, there are no specific directions for the oscillations of the electric and magnetic fields; they vibrate in any randomly oriented plane perpendicular to the direction of propagation. **Polarization** is the attribute that a wave's oscillations do have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus, we can think of the electric field arrows as showing the direction of polarization, as in [link].



An EM wave, such as light, is a transverse wave. The electric $(\vec{\mathbf{E}})$ and magnetic $(\vec{\mathbf{B}})$ fields are perpendicular to the direction of propagation. The direction of polarization of the wave is the direction of the electric field.

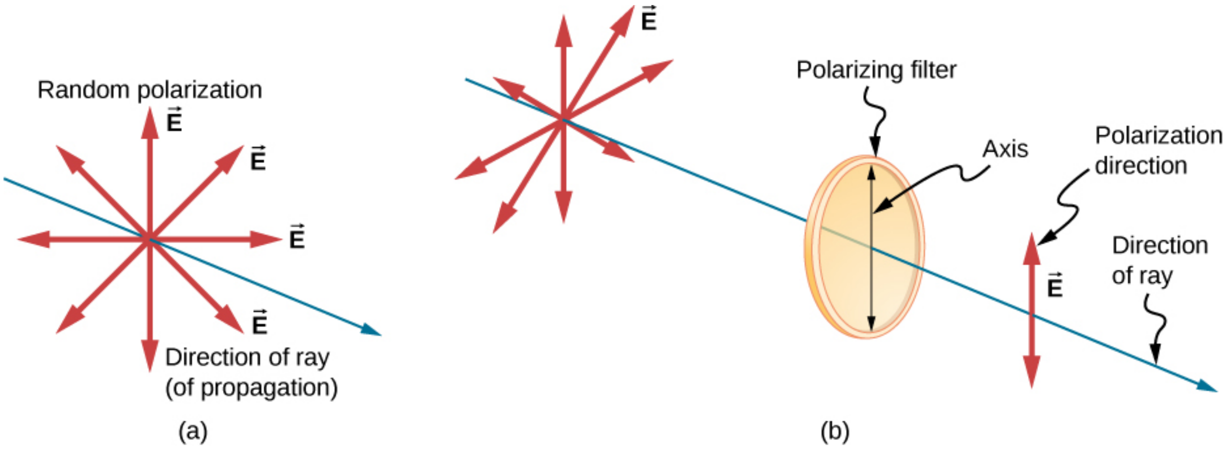
To examine this further, consider the transverse waves in the ropes shown in [link]. The oscillations in one rope are in a vertical plane and are said to be

vertically polarized. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.



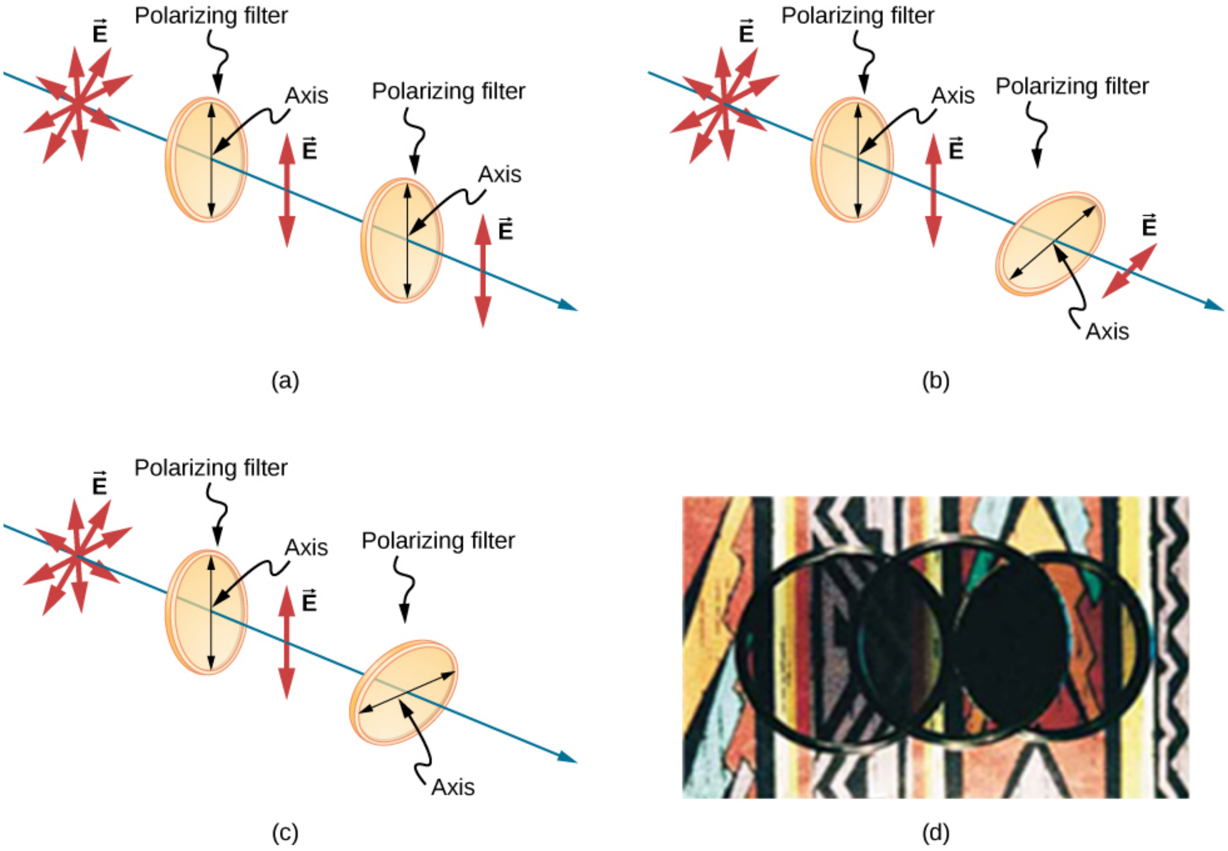
The transverse oscillations in one rope (a) are in a vertical plane, and those in the other rope (b) are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that have the electric fields in random directions ([link](a)). Such light is said to be **unpolarized**, because it is composed of many waves with all possible directions of polarization. Polaroid materials—which were invented by the founder of the Polaroid Corporation, Edwin Land—act as a polarizing slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. If we think of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The axis of a polarizing filter is the direction along which the filter passes the electric field of an EM wave.


Random polarization
$\vec{E}$
$\vec{E}$
$\vec{E}$
$\vec{E}$
Direction of ray
(of propagation)
(a)
$\vec{E}$
Polarizing filter
Axis
Polarization
direction
Direction
of ray
$\vec{E}$
(b)

The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. (a) If the light is unpolarized, the arrows point in all directions. (b) A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

[link] shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second filter. If the second polarizing filter is rotated, only the component of the light parallel to the second filter’s axis is passed. When the axes are perpendicular, no light is passed by the second filter.


$\vec{E}$
Polarizing filter
Axis
$\vec{E}$
Polarizing filter
Axis
$\vec{E}$
(a)
$\vec{E}$
Polarizing filter
Axis
$\vec{E}$
Polarizing filter
$\vec{E}$
Axis
(b)
$\vec{E}$
Polarizing filter
Axis
$\vec{E}$
Polarizing filter
Axis
(c)
(d)

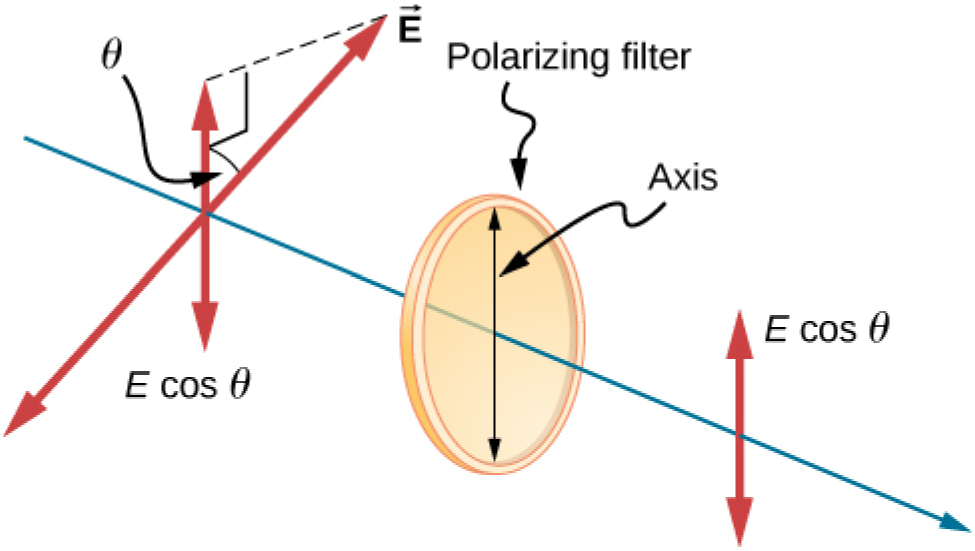
The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second filter is rotated, only part of the light is passed. (c) When the second filter is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit d: modification of work by P.P. Urone)

Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter θ. If the electric field has an amplitude E, then the transmitted part of the wave has an amplitude $E \cos \theta$ ([link]). Since the intensity of a wave is proportional to its amplitude squared, the intensity I of the transmitted wave is related to the incident wave by

Note:
Equation:

$$I = I_0 \cos^2 \theta$$

where I_0 is the intensity of the polarized wave before passing through the filter. This equation is known as **Malus's law**.



θ
$\vec{\mathbf{E}}$
Polarizing filter
Axis
$E \cos \theta$
$E \cos \theta$

A polarizing filter transmits only the component of the wave parallel to its axis, reducing the intensity of any light not polarized parallel to its axis.

Note:
This Open Source Physics animation helps you visualize the electric field vectors as light encounters a polarizing filter. You can rotate the filter—

note that the angle displayed is in radians. You can also rotate the animation for 3D visualization.

Example:

Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by 90.0%?

Strategy

When the intensity is reduced by 90.0%, it is 10.0% or 0.100 times its original value. That is, $I = 0.100\ I_0$. Using this information, the equation $I = I_0 \cos^2 \theta$ can be used to solve for the needed angle.

Solution

Solving the equation $I = I_0 \cos^2 \theta$ for $\cos \theta$ and substituting with the relationship between I and I_0 gives

Equation:

$$\cos \theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100\ I_0}{I_0}} = 0.3162.$$

Solving for θ yields

Equation:

$$\theta = \cos^{-1} 0.3162 = 71.6^\circ.$$

Significance

A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to 10.0% of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that at an angle of 45°, the intensity is reduced to 50% of its original value. Note that 71.6° is 18.4° from reducing the intensity to zero, and that at an angle of 18.4°, the intensity is reduced to 90.0% of its original value, giving evidence of symmetry.

Note:
Exercise:

Problem:

Check Your Understanding Although we did not specify the direction in [link], let’s say the polarizing filter was rotated clockwise by 71.6° to reduce the light intensity by 90.0%. What would be the intensity reduction if the polarizing filter were rotated counterclockwise by 71.6°?

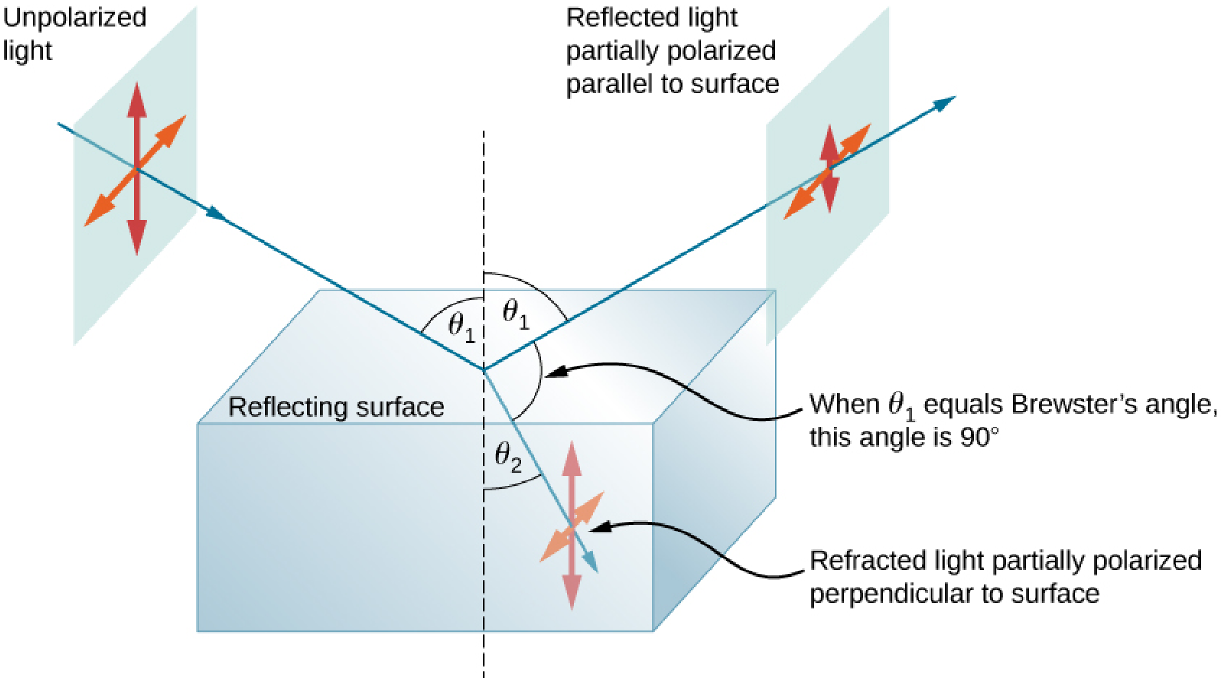
Solution:

also 90.0%

Polarization by Reflection

By now, you can probably guess that polarizing sunglasses cut the glare in reflected light, because that light is polarized. You can check this for yourself by holding polarizing sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

[link] illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so the reflected light is left more horizontally polarized. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization is like an arrow perpendicular to the surface and is more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and is more likely to be reflected. Sunglasses with vertical axes thus block more reflected light than unpolarized light from other sources.


Unpolarized light
Reflected light partially polarized parallel to surface
θ_1
θ_1
Reflecting surface
θ_2
When θ_1 equals Brewster's angle, this angle is 90°
Refracted light partially polarized perpendicular to surface

Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides and bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that reflected light is completely polarized at an angle of reflection θ_{b} given by

Note:
Equation:

$$\tan\theta_{\text{b}} = \frac{n_2}{n_1}$$

where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as **Brewster's law** and θ_b is known as **Brewster's angle**, named after the nineteenth-century Scottish physicist who discovered them.

Note:
This Open Source Physics animation shows incident, reflected, and refracted light as rays and EM waves. Try rotating the animation for 3D visualization and also change the angle of incidence. Near Brewster's angle, the reflected light becomes highly polarized.

Example:
Calculating Polarization by Reflection
(a) At what angle will light traveling in air be completely polarized horizontally when reflected from water? (b) From glass?
Strategy
All we need to solve these problems are the indices of refraction. Air has $n_1 = 1.00$, water has $n_2 = 1.333$, and crown glass has $n'_2 = 1.520$. The equation $\tan\theta_\mathrm{b} = \frac{n_2}{n_1}$ can be directly applied to find θ_b in each case.
Solution

a. Putting the known quantities into the equation
   Equation:

$$\tan\theta_\mathrm{b} = \frac{n_2}{n_1}$$

   gives
   Equation:

$$\tan\theta_\mathrm{b} = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333.$$

Solving for the angle θ_b yields
Equation:

$$\theta_\text{b} = \tan^{-1} 1.333 = 53.1^\circ.$$

b. Similarly, for crown glass and air,
Equation:

$$\tan\theta'_\text{b} = \frac{n'_2}{n_1} = \frac{1.520}{1.00} = 1.52.$$

Thus,
Equation:

$$\theta'_\text{b} = \tan^{-1} 1.52 = 56.7^\circ.$$

Significance
Light reflected at these angles could be completely blocked by a good polarizing filter held with its axis vertical. Brewster's angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light that is not reflected is refracted into these media. Therefore, at an incident angle equal to Brewster's angle, the refracted light is slightly polarized vertically. It is not completely polarized vertically, because only a small fraction of the incident light is reflected, so a significant amount of horizontally polarized light is refracted.

Note:
Exercise:

Problem:

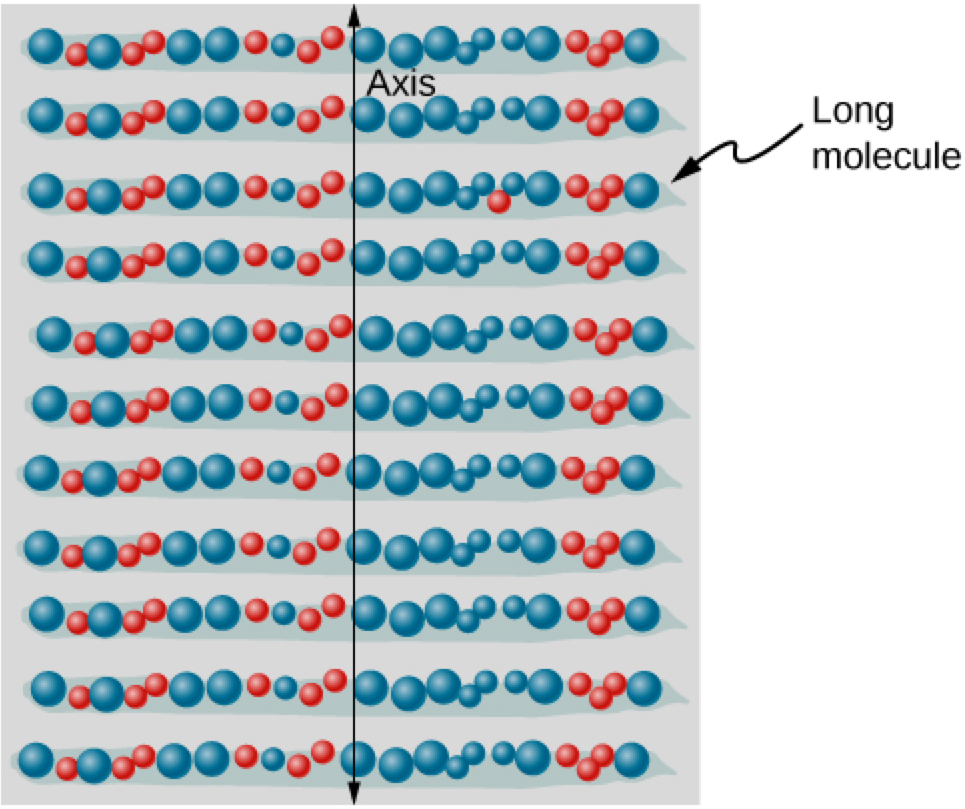
Check Your Understanding What happens at Brewster’s angle if the original incident light is already 100% vertically polarized?

Solution:

There will be only refraction but no reflection.

Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes EM waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis, as shown in [link].



Long molecules are aligned perpendicular to the axis of a polarizing filter. In an EM wave, the component of the electric field perpendicular to these molecules passes through the filter, whereas the component parallel to the molecules is absorbed.

[link] illustrates how the component of the electric field parallel to the long molecules is absorbed. An EM wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons, since electron masses are small. If an electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the field in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component

of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and allow these fields to pass. Thus, the axis of the polarizing filter is perpendicular to the length of the molecule.

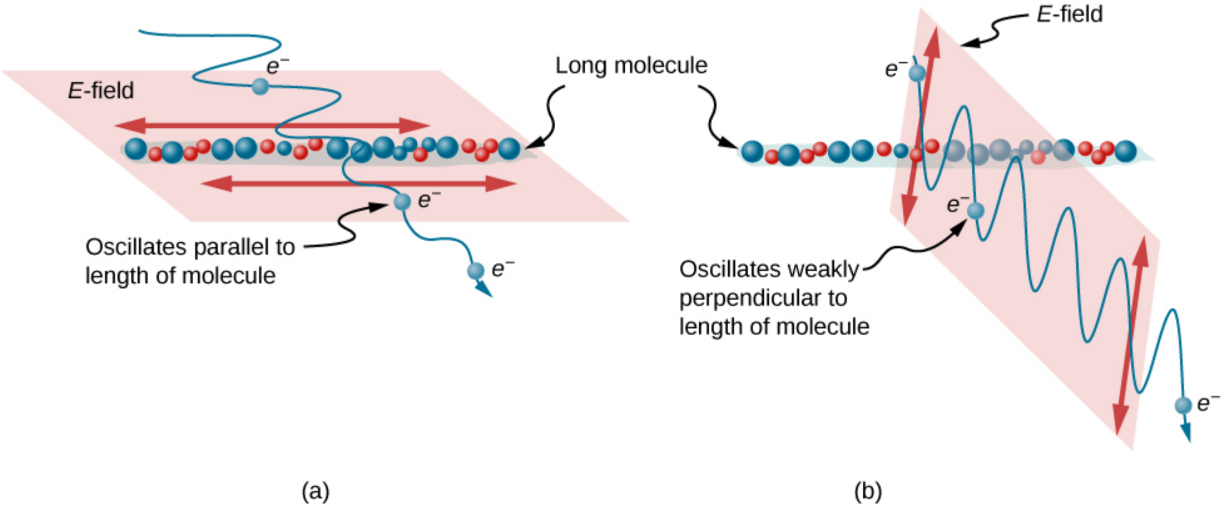
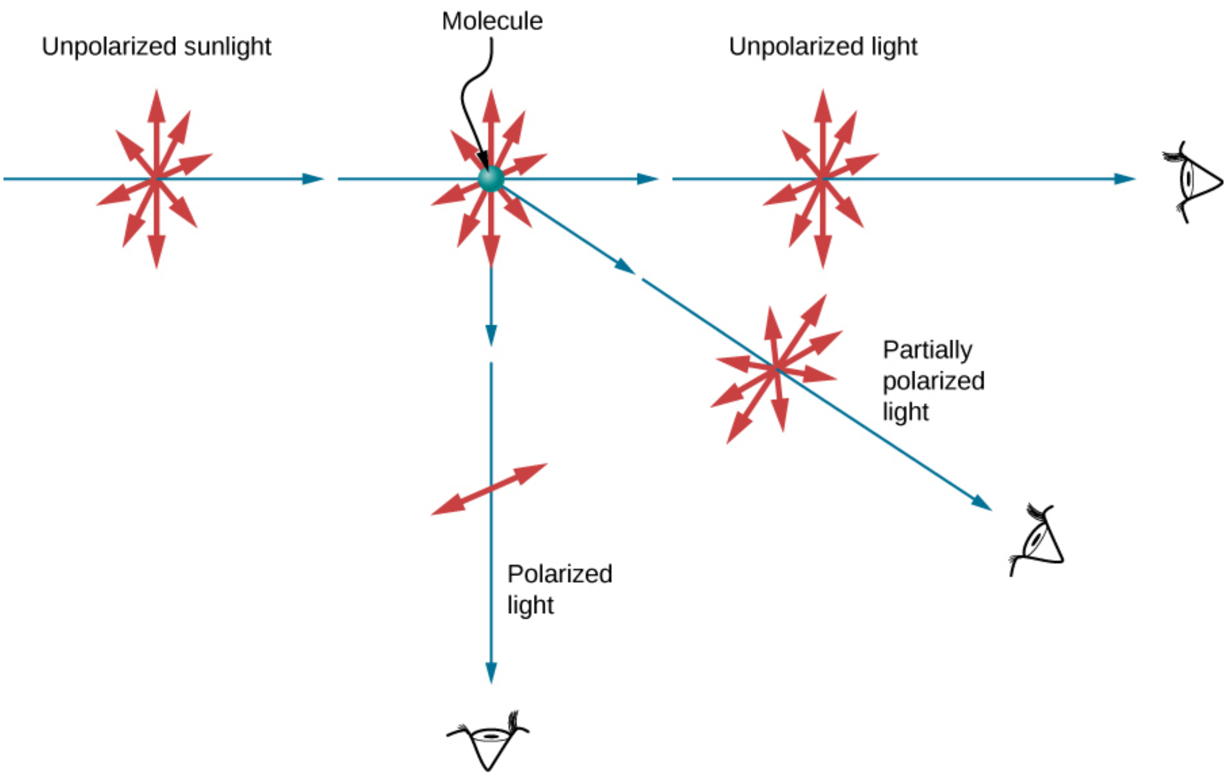

E-field
e⁻
Long molecule
e⁻
Oscillates parallel to length of molecule
e⁻
E-field
e⁻
e⁻
Oscillates weakly perpendicular to length of molecule
e⁻
(a)
(b)

Diagram of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Polarization by Scattering

If you hold your polarizing sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. [link] helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction that it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in the figure, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light is only partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.



Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

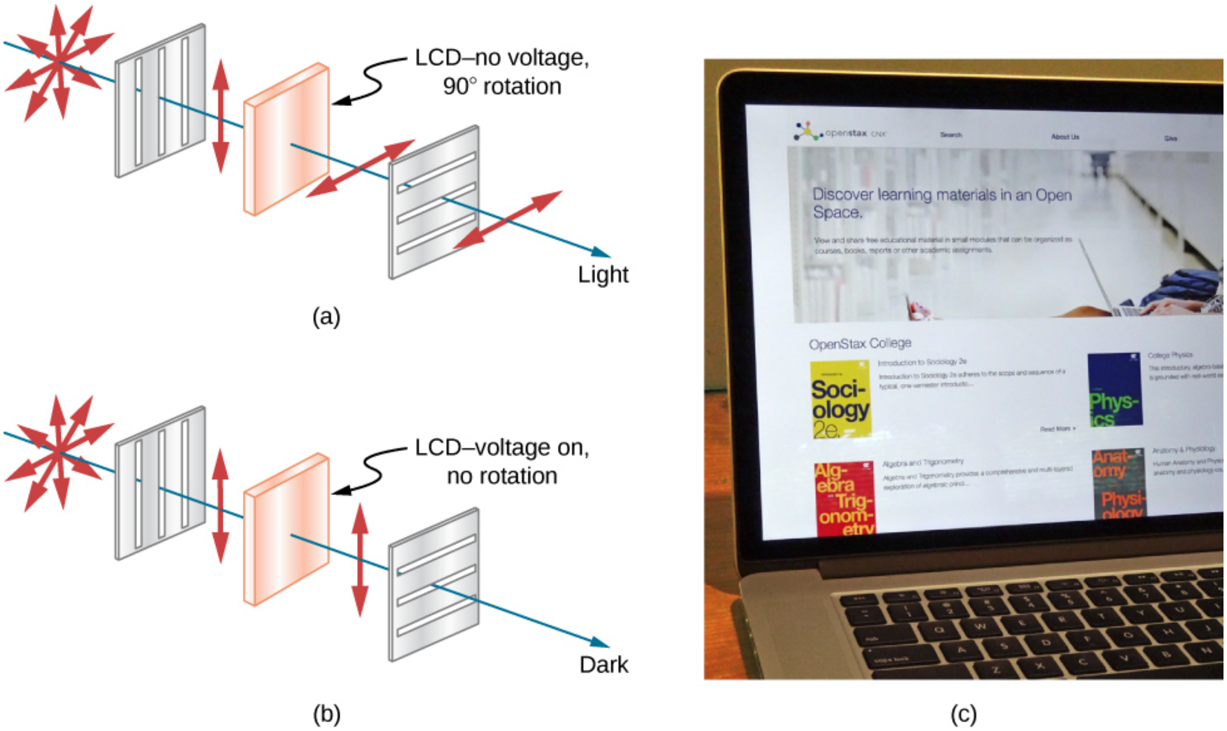
Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

A range of optical effects are used in sunglasses. Besides being polarizing, sunglasses may have colored pigments embedded in them, whereas others use either a nonreflective or reflective coating. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

Liquid Crystals and Other Polarization Effects in Materials

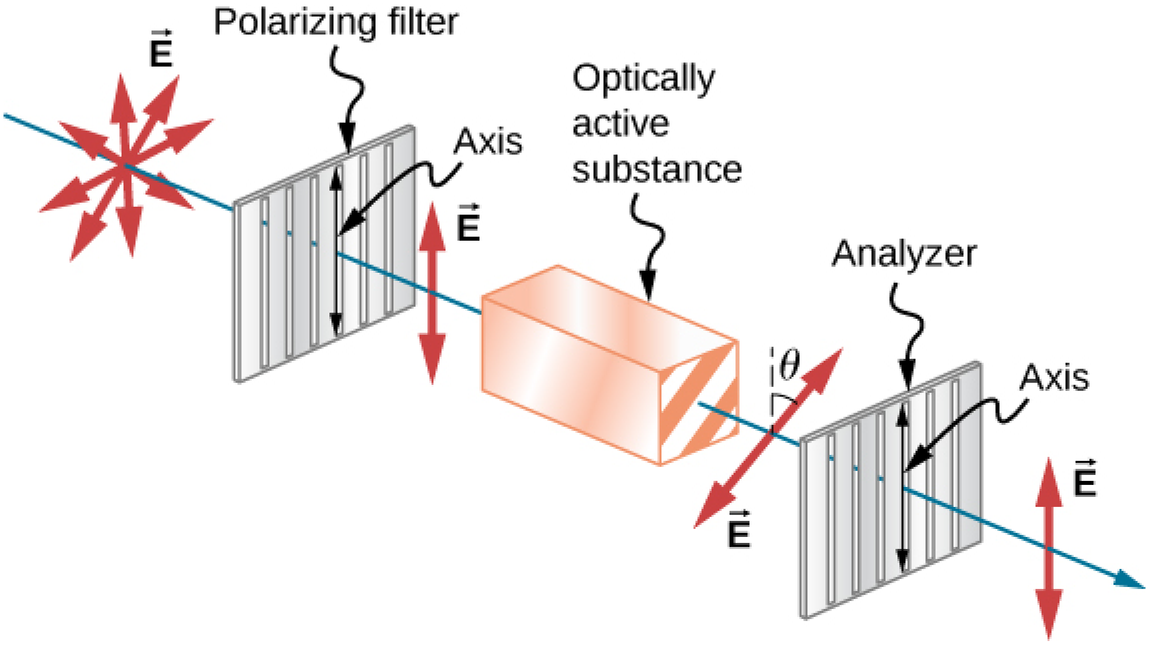
Although you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and many other places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by 90°. Furthermore, this property can be turned off by the application of a voltage, as illustrated in [link]. It is possible to manipulate this characteristic quickly and in small, well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, a large light is generated at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in [link](a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. We can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.



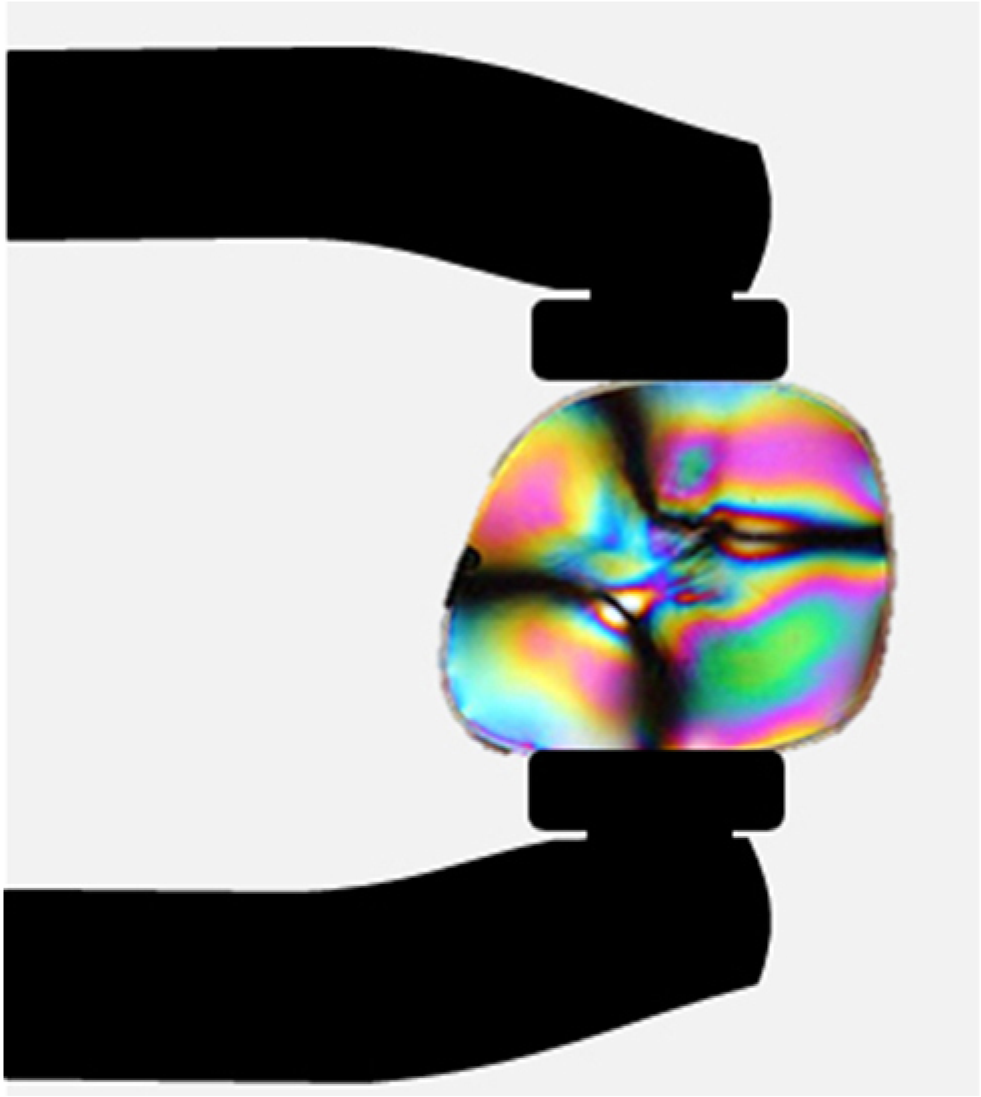
(a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the direction of the original polarization. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit c: modification of work by Jane Whitney)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen ([link]). In addition to depending on the type of substance, the amount and direction of rotation depend on several other factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetrical shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.


Polarizing filter
$\vec{E}$
Axis
$\vec{E}$
Optically active substance
Analyzer
θ
Axis
$\vec{E}$
$\vec{E}$

Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

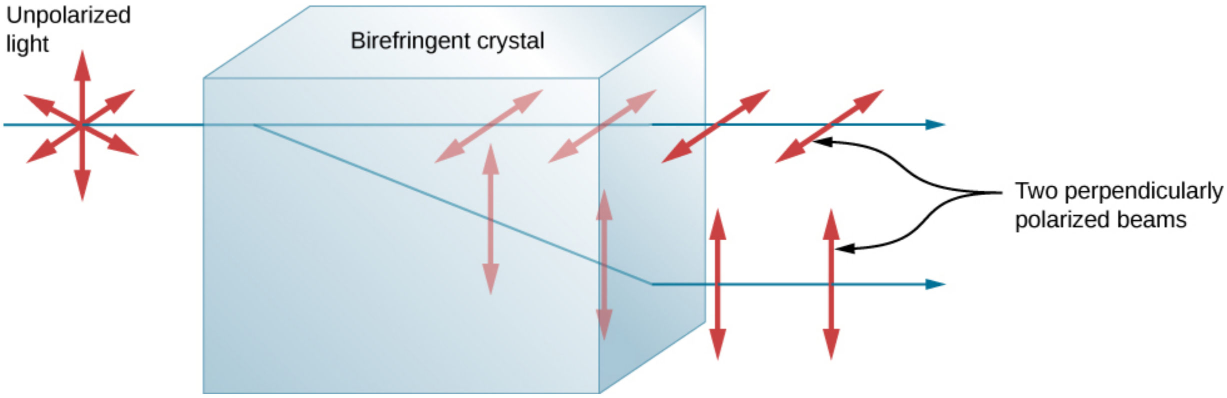
Glass and plastic become optically active when stressed: the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in [link]. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.



Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: “Infopro”/Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two polarized beams. This occurs because the crystal has one value for the index of refraction of polarized light but a different value for the index of refraction of light polarized in the perpendicular direction, so that each component has its own angle of refraction. Such crystals are said to be

birefringent, and, when aligned properly, two perpendicularly polarized beams will emerge from the crystal ([link]). Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work.



Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two with two different values of index of refraction.

Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave. The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Unpolarized light can be polarized by passing it through a polarizing filter or other polarizing material. The process of polarizing light decreases its intensity by a factor of 2.

- The intensity, *I*, of polarized light after passing through a polarizing filter is $I = I_0 \cos^2 \theta$, where I_0 is the incident intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster’s law states that reflected light is completely polarized at the angle of reflection θ_b, known as Brewster’s angle.
- Polarization can also be produced by scattering.
- Several types of optically active substances rotate the direction of polarization of light passing through them.

Key Equations

Speed of light	$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$
Index of refraction	$n = \frac{c}{v}$
Law of reflection	$\theta_\text{r} = \theta_\text{i}$
Law of refraction (Snell’s law)	$n_1 \sin \theta_1 = n_2 \sin \theta_2$
Critical angle	$\theta_\text{c} = \sin^{-1} \left(\frac{n_2}{n_1} \right) \text{ for } n_1 > n_2$
Malus’s law	$I = I_0 \cos^2 \theta$
Brewster’s law	$\tan \theta_\text{b} = \frac{n_2}{n_1}$

Conceptual Questions

Exercise:

Problem: Can a sound wave in air be polarized? Explain.

Solution:

No. Sound waves are not transverse waves.

Exercise:

Problem:

No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?

Exercise:

Problem:

Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.

Solution:

Energy is absorbed into the filters.

Exercise:

Problem:

When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $\frac{1}{\lambda}$. Does this mean there is more scattering for small λ than large λ? How does this relate to the fact that the sky is blue?

Exercise:

Problem:

Using the information given in the preceding question, explain why sunsets are red.

Solution:

Sunsets are viewed with light traveling straight from the Sun toward us. When blue light is scattered out of this path, the remaining red light dominates the overall appearance of the setting Sun.

Exercise:

Problem:

When light is reflected at Brewster’s angle from a smooth surface, it is 100% polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be 100%?

Exercise:

Problem:

If you lie on a beach looking at the water with your head tipped slightly sideways, your polarized sunglasses do not work very well. Why not?

Solution:

The axis of polarization for the sunglasses has been rotated 90°.

Problems

Exercise:

Problem:

What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?

Exercise:

Problem:

The angle between the axes of two polarizing filters is 45.0°. By how much does the second filter reduce the intensity of the light coming through the first?

Solution:

0.500

Exercise:

Problem:

Two polarizing sheets P_1 and P_2 are placed together with their transmission axes oriented at an angle θ to each other. What is θ when only 25% of the maximum transmitted light intensity passes through them?

Exercise:

Problem:

Suppose that in the preceding problem the light incident on P_1 is unpolarized. At the determined value of θ, what fraction of the incident light passes through the combination?

Solution:

0.125 or 1/8

Exercise:

Problem:

If you have completely polarized light of intensity $150\ \text{W/m}^2$, what will its intensity be after passing through a polarizing filter with its axis at an 89.0° angle to the light's polarization direction?

Exercise:

Problem:

What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity $1.00\ \text{kW/m}^2$ to reduce the intensity to $10.0\ \text{W/m}^2$?

Solution:

84.3°

Exercise:

Problem:

At the end of [link], it was stated that the intensity of polarized light is reduced to 90.0% of its original value by passing through a polarizing filter with its axis at an angle of 18.4° to the direction of polarization. Verify this statement.

Exercise:

Problem:

Show that if you have three polarizing filters, with the second at an angle of 45.0° to the first and the third at an angle of 90.0° to the first, the intensity of light passed by the first will be reduced to 25.0% of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)

Solution:

$0.250\ I_0$

Exercise:

Problem:

Three polarizing sheets are placed together such that the transmission axis of the second sheet is oriented at 25.0° to the axis of the first, whereas the transmission axis of the third sheet is oriented at 40.0° (in the same sense) to the axis of the first. What fraction of the intensity of an incident unpolarized beam is transmitted by the combination?

Exercise:

Problem:

In order to rotate the polarization axis of a beam of linearly polarized light by 90.0°, a student places sheets P_1 and P_2 with their transmission axes at 45.0° and 90.0°, respectively, to the beam's axis of polarization. (a) What fraction of the incident light passes through P_1 and (b) through the combination? (c) Repeat your calculations for part (b) for transmission-axis angles of 30.0° and 90.0°, respectively.

Solution:

a. 0.500; b. 0.250; c. 0.187

Exercise:

Problem:

It is found that when light traveling in water falls on a plastic block, Brewster's angle is 50.0°. What is the refractive index of the plastic?

Exercise:

Problem:

At what angle will light reflected from diamond be completely polarized?

Solution:

67.54°

Exercise:

Problem:

What is Brewster’s angle for light traveling in water that is reflected from crown glass?

Exercise:

Problem:

A scuba diver sees light reflected from the water’s surface. At what angle relative to the water’s surface will this light be completely polarized?

Solution:

53.1°

Additional Problems

Exercise:

Problem:

From his measurements, Roemer estimated that it took 22 min for light to travel a distance equal to the diameter of Earth’s orbit around the Sun. (a) Use this estimate along with the known diameter of Earth’s orbit to obtain a rough value of the speed of light. (b) Light actually takes 16.5 min to travel this distance. Use this time to calculate the speed of light.

Exercise:

Problem:

Cornu performed Fizeau’s measurement of the speed of light using a wheel of diameter 4.00 cm that contained 180 teeth. The distance from the wheel to the mirror was 22.9 km. Assuming he measured the speed of light accurately, what was the angular velocity of the wheel?

Solution:

114 radian/s

Exercise:

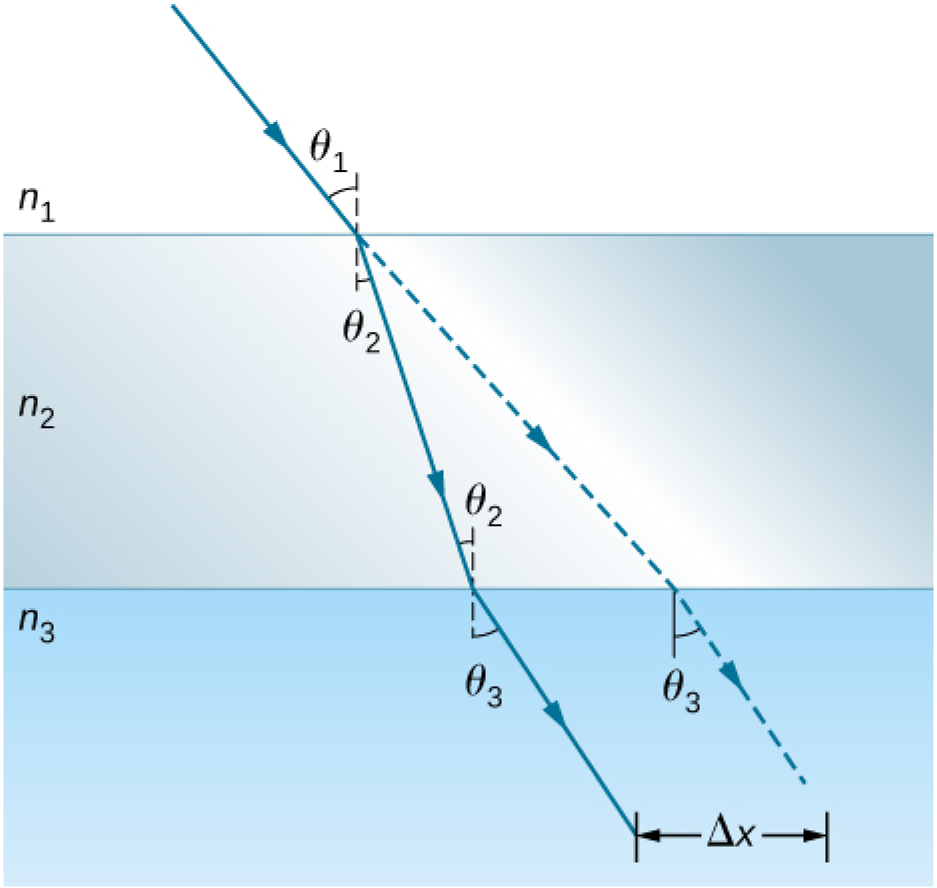
Problem:

Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of 45.0°, and you observe the angle of refraction to be 40.3°. What is the index of refraction of the substance and its likely identity?

Exercise:

Problem:

Shown below is a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is 40.0° and the glass is 1.00 cm thick.


θ_1
n_1
θ_2
n_2
θ_2
n_3
θ_3
θ_3
Δx

Solution:

3.72 mm

Exercise:

Problem:

Considering the previous problem, show that θ_3 is the same as it would be if the second medium were not present.

Exercise:

Problem:

At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?

Solution:

41.2°

Exercise:

Problem:

Light reflected at 55.6° from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?

Exercise:

Problem:

(a) Light reflected at 62.5° from a gemstone in a ring is completely polarized. Can the gem be a diamond? (b) At what angle would the light be completely polarized if the gem was in water?

Solution:

a. 1.92. The gem is not a diamond (it is zircon). b. 55.2°

Exercise:

Problem:

If θ_{b} is Brewster's angle for light reflected from the top of an interface between two substances, and θ'_{b} is Brewster's angle for light reflected from below, prove that $\theta_{\mathrm{b}} + \theta'_{\mathrm{b}} = 90.0^{\circ}$.

Exercise:

Problem:

Unreasonable results Suppose light travels from water to another substance, with an angle of incidence of 10.0° and an angle of refraction of 14.9°. (a) What is the index of refraction of the other substance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. 0.898; b. We cannot have $n < 1.00$, since this would imply a speed greater than c. c. The refracted angle is too big relative to the angle of incidence.

Exercise:

Problem:

Unreasonable results Light traveling from water to a gemstone strikes the surface at an angle of 80.0° and has an angle of refraction of 15.2°. (a) What is the speed of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem:

If a polarizing filter reduces the intensity of polarized light to 50.0% of its original value, by how much are the electric and magnetic fields reduced?

Solution:

$0.707\ B_1$

Exercise:

Problem:

Suppose you put on two pairs of polarizing sunglasses with their axes at an angle of 15.0°. How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.

Exercise:

Problem:

(a) On a day when the intensity of sunlight is $1.00\,\mathrm{kW/m^2}$, a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of 20.0°. Assuming the sunlight is unpolarized and the polarizers are 100% efficient, what is the initial rate of heating of the water in $^\circ\mathrm{C/s}$, assuming it is 80.0% absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water. (b) Do the polarizing filters get hot? Explain.

Solution:

a. $1.69\ \times\ 10^{-2}\,^\circ\mathrm{C/s}$; b. yes

Challenge Problems

Exercise:

Problem:

Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ.

Exercise:

Problem:

Consider sunlight entering Earth’s atmosphere at sunrise and sunset—that is, at a 90.0° incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.

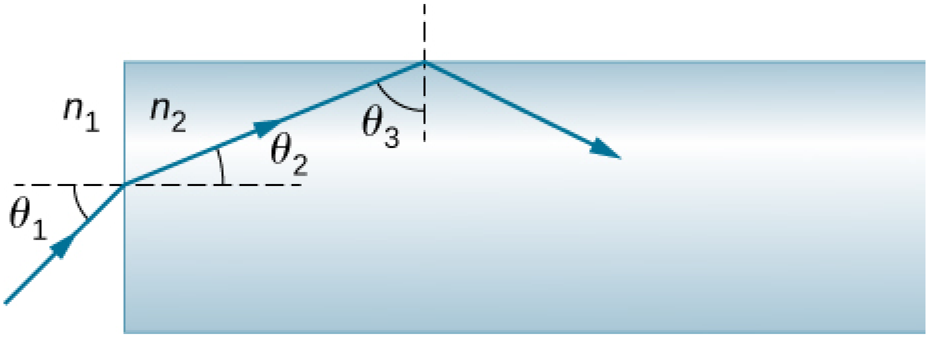
Solution:

First part: 88.6°. The remainder depends on the complexity of the solution the reader constructs.

Exercise:

Problem:

A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown below. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.



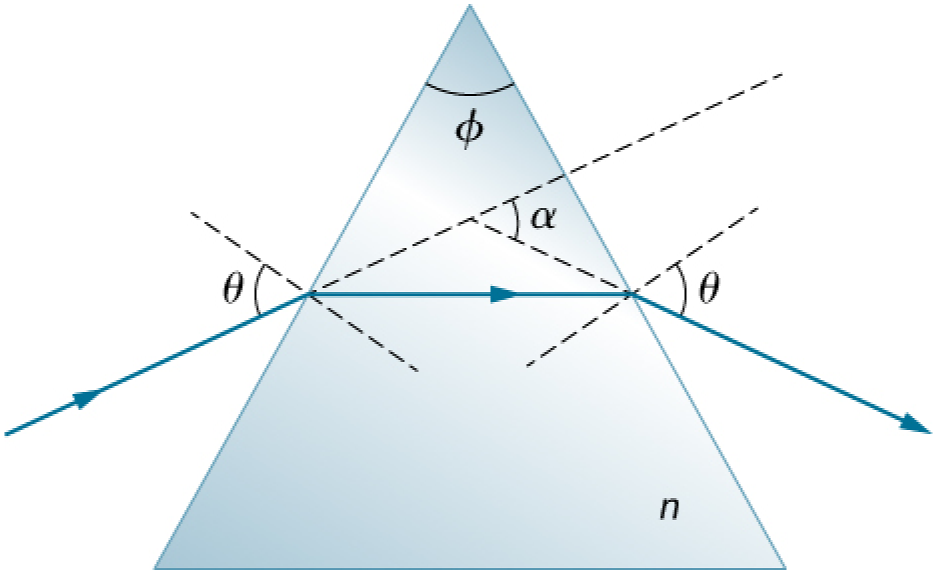
Exercise:

Problem:

A light ray falls on the left face of a prism (see below) at the angle of incidence θ for which the emerging beam has an angle of refraction θ at the right face. Show that the index of refraction *n* of the glass prism is given by

$$n = \frac{\sin \frac{1}{2}(\alpha + \phi)}{\sin \frac{1}{2}\phi}$$

where ϕ is the vertex angle of the prism and α is the angle through which the beam has been deviated. If $\alpha = 37.0^\circ$ and the base angles of the prism are each 50.0°, what is *n*?



Solution:

proof; 1.33

Exercise:

Problem:

If the apex angle ϕ in the previous problem is 20.0° and $n = 1.50$, what is the value of α?

Exercise:

Problem:

The light incident on polarizing sheet P_1 is linearly polarized at an angle of 30.0° with respect to the transmission axis of P_1. Sheet P_2 is placed so that its axis is parallel to the polarization axis of the incident light, that is, also at 30.0° with respect to P_1. (a) What fraction of the incident light passes through P_1? (b) What fraction of the incident light is passed by the combination? (c) By rotating P_2, a maximum in transmitted intensity is obtained. What is the ratio of this maximum intensity to the intensity of transmitted light when P_2 is at 30.0° with respect to P_1?

Solution:

a. 0.750; b. 0.563; c. 1.33

Exercise:

Problem:

Prove that if I is the intensity of light transmitted by two polarizing filters with axes at an angle θ and $I\prime$ is the intensity when the axes are at an angle $90.0^\circ - \theta$, then $I + I' = I_0$, the original intensity. (*Hint:* Use the trigonometric identities $\cos 90.0^\circ - \theta = \sin \theta$ and $\cos^2 \theta + \sin^2 \theta = 1.$)

Glossary

birefringent
refers to crystals that split an unpolarized beam of light into two beams

Brewster’s angle
angle of incidence at which the reflected light is completely polarized

Brewster’s law
$\tan \theta_\mathrm{b} = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light

direction of polarization
direction parallel to the electric field for EM waves

horizontally polarized
oscillations are in a horizontal plane

Malus’s law
where I_0 is the intensity of the polarized wave before passing through the filter

optically active
substances that rotate the plane of polarization of light passing through them

polarization
attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized
refers to waves having the electric and magnetic field oscillations in a definite direction

unpolarized
refers to waves that are randomly polarized

vertically polarized
oscillations are in a vertical plane

Introduction class="introduction"

Cloud Gate is a public sculpture by Anish Kapoor located in Millennium Park in Chicago. Its stainless steel plates reflect and distort images around it, including the Chicago skyline. Dedicated in 2006, it has become a popular tourist attraction, illustrating how art can use the principles of physical optics to startle and entertain. (credit: modificatio

n of work
by Dhilung
Kirat)

This chapter introduces the major ideas of geometric optics, which describe the formation of images due to reflection and refraction. It is called “geometric” optics because the images can be characterized using geometric constructions, such as ray diagrams. We have seen that visible light is an electromagnetic wave; however, its wave nature becomes evident only when light interacts with objects with dimensions comparable to the wavelength (about 500 nm for visible light). Therefore, the laws of geometric optics only apply to light interacting with objects much larger than the wavelength of the light.

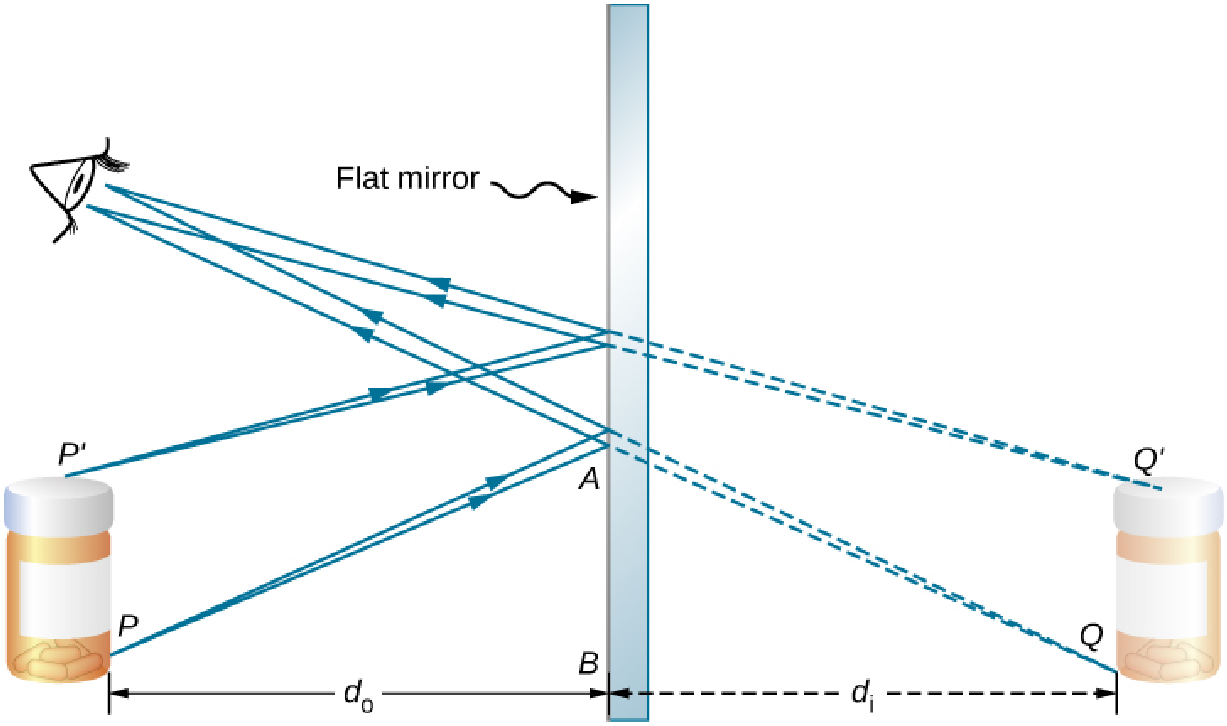
Images Formed by Plane Mirrors

By the end of this section, you will be able to:

- Describe how an image is formed by a plane mirror.
- Distinguish between real and virtual images.
- Find the location and characterize the orientation of an image created by a plane mirror.

You only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in a **plane mirror** are the same size as the object, are located behind the mirror, and are oriented in the same direction as the object (i.e., “upright”).

To understand how this happens, consider [link]. Two rays emerge from point P, strike the mirror, and reflect into the observer’s eye. Note that we use the law of reflection to construct the reflected rays. If the reflected rays are extended backward behind the mirror (see dashed lines in [link]), they seem to originate from point Q. This is where the image of point P is located. If we repeat this process for point P', we obtain its image at point Q'. You should convince yourself by using basic geometry that the image height (the distance from Q to Q') is the same as the object height (the distance from P to P'). By forming images of all points of the object, we obtain an upright image of the object behind the mirror.



Two light rays originating from point P on an object are reflected by a flat mirror into the eye of an observer. The reflected rays are obtained by using the law of reflection. Extending these reflected rays backward, they seem to come from point Q behind the mirror, which is where the virtual image is located. Repeating this process for point P' gives the image point Q'. The image height is thus the same as the object height, the image is upright, and the object distance d_o is the same as the image distance d_i. (credit: modification of work by Kevin Dufendach)

Notice that the reflected rays appear to the observer to come directly from the image behind the mirror. In reality, these rays come from the points on the mirror where they are reflected. The image behind the mirror is called a **virtual image** because it cannot be projected onto a screen—the rays only appear to originate from a common point behind the mirror. If you walk behind the mirror, you cannot see the image, because the rays do not go there. However, in front of the mirror, the rays behave exactly as if they come from behind the mirror, so that is where the virtual image is located.

Later in this chapter, we discuss real images; a **real image** can be projected onto a screen because the rays physically go through the image. You can certainly see both real and virtual images. The difference is that a virtual image cannot be projected onto a screen, whereas a real image can.

Locating an Image in a Plane Mirror

The law of reflection tells us that the angle of incidence is the same as the angle of reflection. Applying this to triangles *PAB* and *QAB* in [link] and using basic geometry shows that they are congruent triangles. This means that the distance *PB* from the object to the mirror is the same as the distance *BQ* from the mirror to the image. The **object distance** (denoted d_o) is the distance from the mirror to the object (or, more generally, from the center of the optical element that creates its image). Similarly, the **image distance** (denoted d_i) is the distance from the mirror to the image (or, more generally, from the center of the optical element that creates it). If we measure distances from the mirror, then the object and image are in opposite directions, so for a plane mirror, the object and image distances should have the opposite signs:

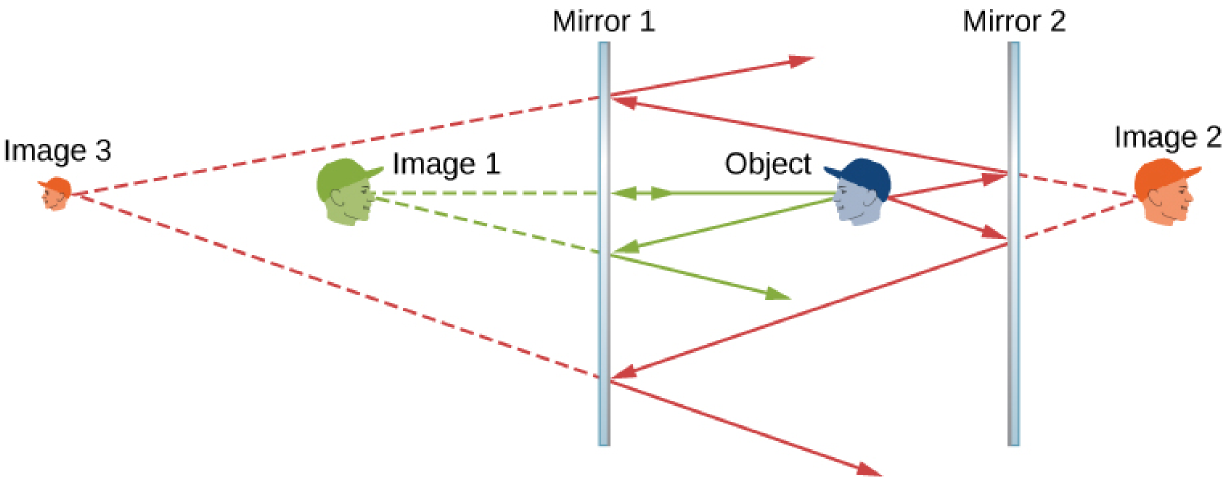
Note:
Equation:

$$d_{\mathrm{o}} = -d_{\mathrm{i}}.$$

An extended object such as the container in [link] can be treated as a collection of points, and we can apply the method above to locate the image of each point on the extended object, thus forming the extended image.

Multiple Images

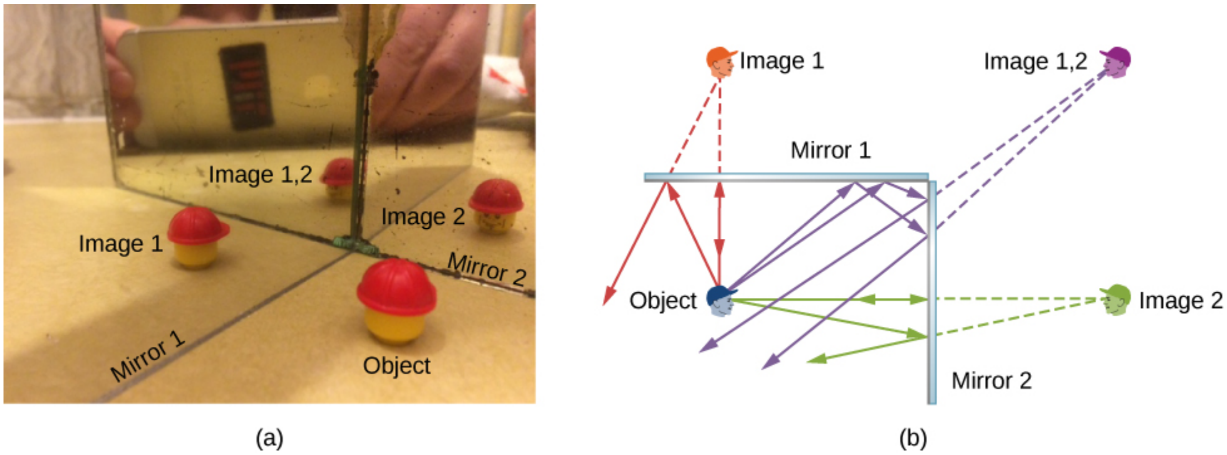
If an object is situated in front of two mirrors, you may see images in both mirrors. In addition, the image in the first mirror may act as an object for the second mirror, so the second mirror may form an image of the image. If the mirrors are placed parallel to each other and the object is placed at a point other than the midpoint between them, then this process of image-of-an-image continues without end, as you may have noticed when standing in a hallway with mirrors on each side. This is shown in [link], which shows three images produced by the blue object. Notice that each reflection reverses front and back, just like pulling a right-hand glove inside out produces a left-hand glove (this is why a reflection of your right hand is a left hand). Thus, the fronts and backs of images 1 and 2 are both inverted with respect to the object, and the front and back of image 3 is inverted with respect to image 2, which is the object for image 3.



Two parallel mirrors can produce, in theory, an infinite number of images of an object placed off center between the mirrors. Three of these images are shown here. The front and back of each image is inverted with respect to its object. Note that the colors are only to identify the images. For normal mirrors, the color of an image is essentially the same as that of its object.

You may have noticed that image 3 is smaller than the object, whereas images 1 and 2 are the same size as the object. The ratio of the image height with respect to the object height is called **magnification**. More will be said about magnification in the next section.

Infinite reflections may terminate. For instance, two mirrors at right angles form three images, as shown in part (a) of [link]. Images 1 and 2 result from rays that reflect from only a single mirror, but image 1,2 is formed by rays that reflect from both mirrors. This is shown in the ray-tracing diagram in part (b) of [link]. To find image 1,2, you have to look behind the corner of the two mirrors.



Two mirrors can produce multiple images. (a) Three images of a plastic head are visible in the two mirrors at a right angle. (b) A single object reflecting from two mirrors at a right angle can produce three images, as shown by the green, purple, and red images.

Summary

- A plane mirror always forms a virtual image (behind the mirror).
- The image and object are the same distance from a flat mirror, the image size is the same as the object size, and the image is upright.

Conceptual Questions

Exercise:

Problem:

What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?

Solution:

Virtual image cannot be projected on a screen. You cannot distinguish a real image from a virtual image simply by judging from the image perceived with your eye.

Exercise:

Problem: Can you see a virtual image? Explain your response.

Exercise:

Problem: Can you photograph a virtual image?

Solution:

Yes, you can photograph a virtual image. For example, if you photograph your reflection from a plane mirror, you get a photograph of a virtual image. The camera focuses the light that enters its lens to form an image; whether the source of the light is a real object or a reflection from mirror (i.e., a virtual image) does not matter.

Exercise:

Problem: Can you project a virtual image onto a screen?

Exercise:

Problem: Is it necessary to project a real image onto a screen to see it?

Solution:

No, you can see the real image the same way you can see the virtual image. The retina of your eye effectively serves as a screen.

Exercise:

Problem:

Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?

Exercise:

Problem:

If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.

Solution:

The mirror should be half your size and its top edge should be at the level of your eyes. The size does not depend on your distance from the mirror.

Problems

Exercise:

Problem:

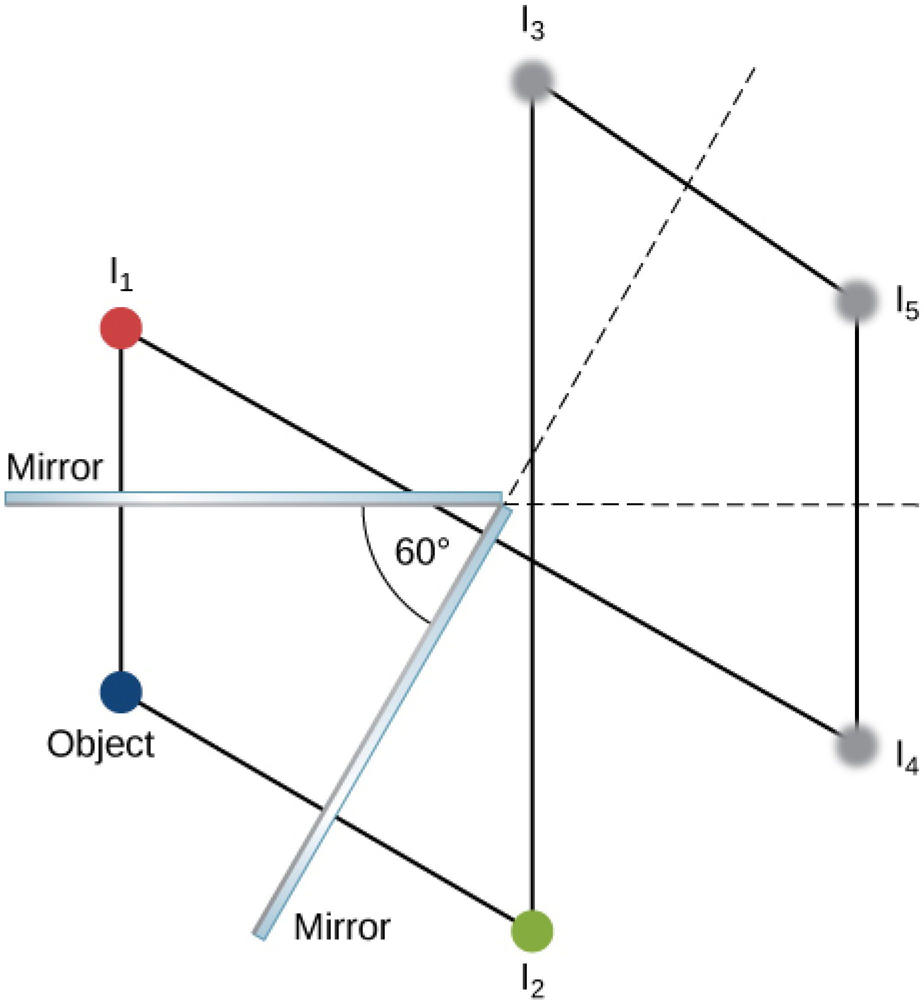
Consider a pair of flat mirrors that are positioned so that they form an angle of 120°. An object is placed on the bisector between the mirrors. Construct a ray diagram as in [link] to show how many images are formed.

Exercise:

Problem:

Consider a pair of flat mirrors that are positioned so that they form an angle of 60°. An object is placed on the bisector between the mirrors. Construct a ray diagram as in [link] to show how many images are formed.

Solution:


I3
I1
I5
Mirror
60°
Object
I4
Mirror
I2

Exercise:

Problem:

By using more than one flat mirror, construct a ray diagram showing how to create an inverted image.

Glossary

plane mirror
plane (flat) reflecting surface

image distance
distance of the image from the central axis of the optical element that produces the image

magnification
ratio of image size to object size

object distance
distance of the object from the central axis of the optical element that produces its image

real image
image that can be projected onto a screen because the rays physically go through the image

virtual image
image that cannot be projected on a screen because the rays do not physically go through the image, they only appear to originate from the image

Spherical Mirrors

By the end of this section, you will be able to:

- Describe image formation by spherical mirrors.
- Use ray diagrams and the mirror equation to calculate the properties of an image in a spherical mirror.

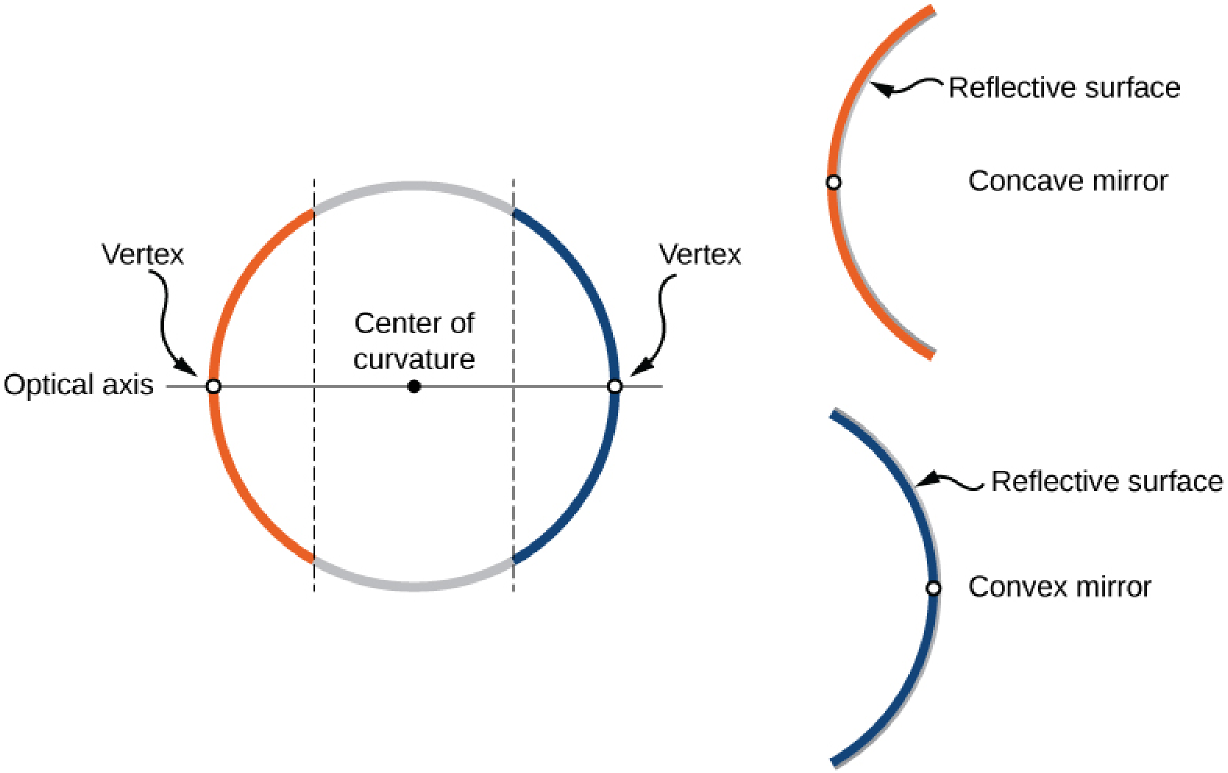
The image in a plane mirror has the same size as the object, is upright, and is the same distance behind the mirror as the object is in front of the mirror. A **curved mirror**, on the other hand, can form images that may be larger or smaller than the object and may form either in front of the mirror or behind it. In general, any curved surface will form an image, although some images make be so distorted as to be unrecognizable (think of fun house mirrors).

Because curved mirrors can create such a rich variety of images, they are used in many optical devices that find many uses. We will concentrate on spherical mirrors for the most part, because they are easier to manufacture than mirrors such as parabolic mirrors and so are more common.

Curved Mirrors

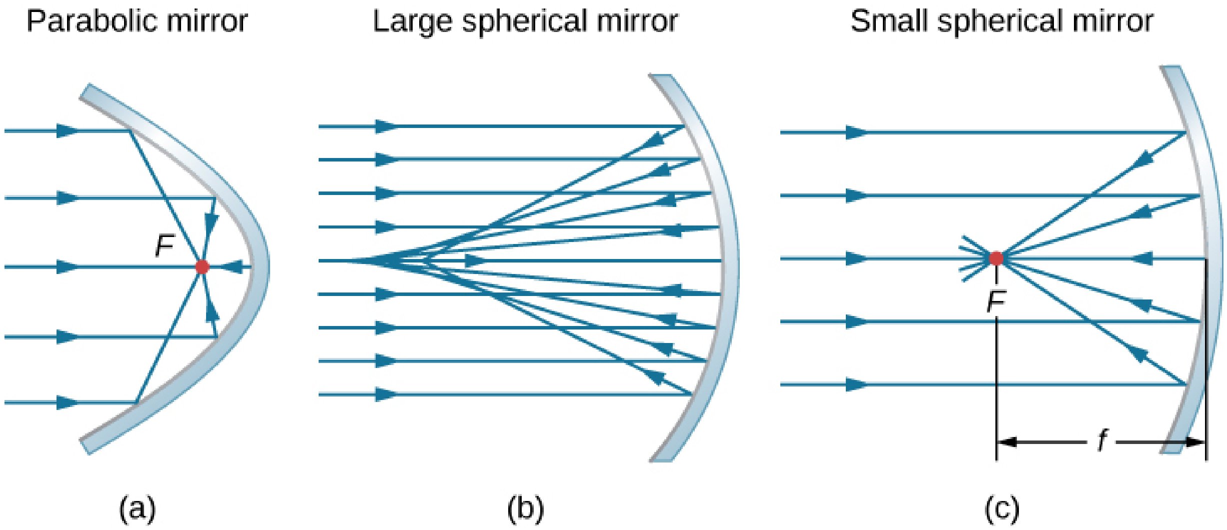
We can define two general types of spherical mirrors. If the reflecting surface is the outer side of the sphere, the mirror is called a **convex mirror**. If the inside surface is the reflecting surface, it is called a **concave mirror**.

Symmetry is one of the major hallmarks of many optical devices, including mirrors and lenses. The symmetry axis of such optical elements is often called the principal axis or **optical axis**. For a spherical mirror, the optical axis passes through the mirror’s center of curvature and the mirror’s **vertex**, as shown in [link].



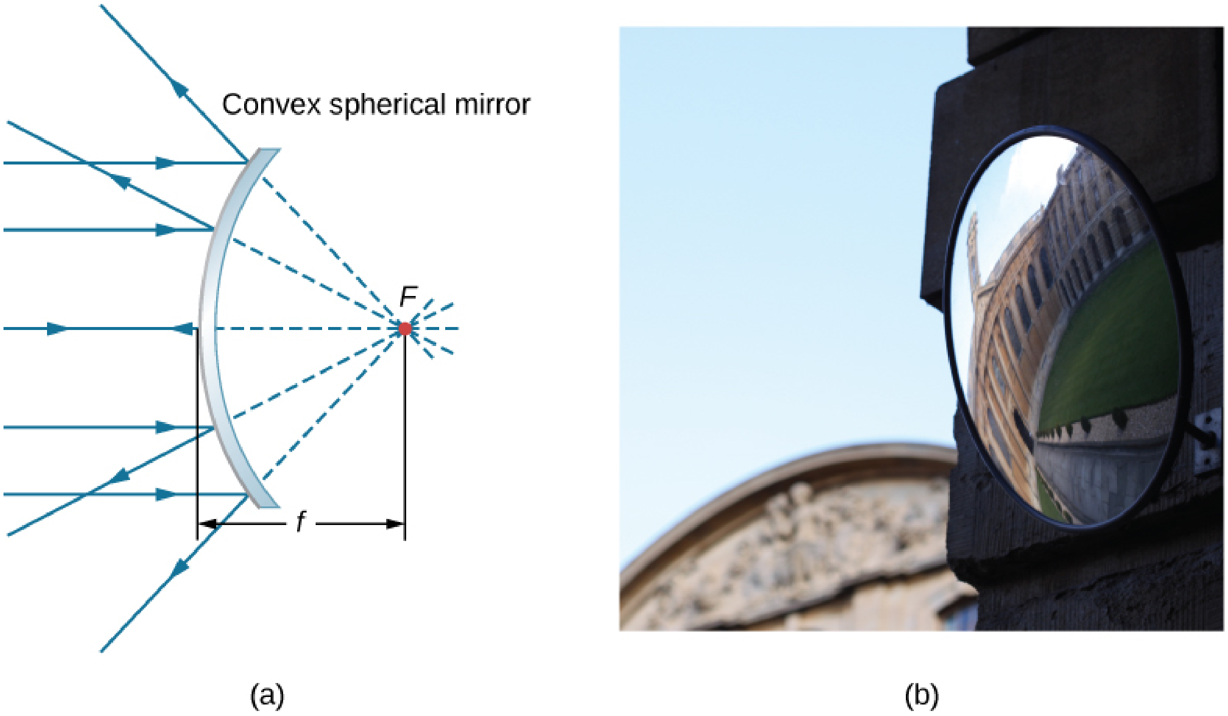
A spherical mirror is formed by cutting out a piece of a sphere and silvering either the inside or outside surface. A concave mirror has silvering on the interior surface (think “cave”), and a convex mirror has silvering on the exterior surface.

Consider rays that are parallel to the optical axis of a parabolic mirror, as shown in part (a) of [link]. Following the law of reflection, these rays are reflected so that they converge at a point, called the **focal point**. Part (b) of this figure shows a spherical mirror that is large compared with its radius of curvature. For this mirror, the reflected rays do not cross at the same point, so the mirror does not have a well-defined focal point. This is called spherical aberration and results in a blurred image of an extended object. Part (c) shows a spherical mirror that is small compared to its radius of curvature. This mirror is a good approximation of a parabolic mirror, so rays that arrive parallel to the optical axis are reflected to a well-defined focal point. The distance along the optical axis from the mirror to the focal point is called the **focal length** of the mirror.



(a) Parallel rays reflected from a parabolic mirror cross at a single point called the focal point *F*. (b) Parallel rays reflected from a large spherical mirror do not cross at a common point. (c) If a spherical mirror is small compared with its radius of curvature, it better approximates the central part of a parabolic mirror, so parallel rays essentially cross at a common point. The distance along the optical axis from the mirror to the focal point is the focal length *f* of the mirror.

A convex spherical mirror also has a focal point, as shown in [link]. Incident rays parallel to the optical axis are reflected from the mirror and seem to originate from point *F* at focal length *f* behind the mirror. Thus, the focal point is virtual because no real rays actually pass through it; they only appear to originate from it.

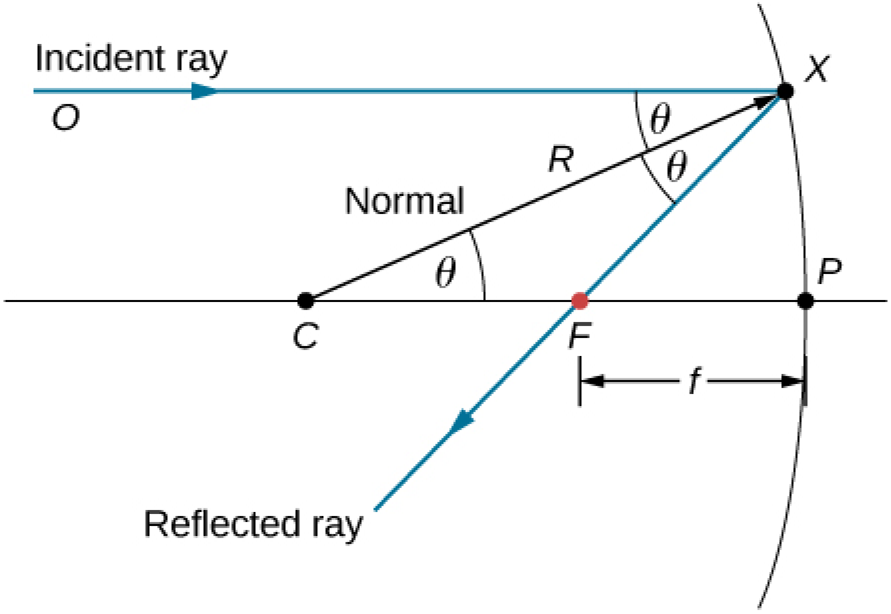


(a) Rays reflected by a convex spherical mirror: Incident rays of light parallel to the optical axis are reflected from a convex spherical mirror and seem to originate from a well-defined focal point at focal distance *f* on the opposite side of the mirror. The focal point is virtual because no real rays pass through it. (b) Photograph of a virtual image formed by a convex mirror. (credit b: modification of work by Jenny Downing)

How does the focal length of a mirror relate to the mirror's radius of curvature? [link] shows a single ray that is reflected by a spherical concave mirror. The incident ray is parallel to the optical axis. The point at which the reflected ray crosses the optical axis is the focal point. Note that all incident rays that are parallel to the optical axis are reflected through the focal point—we only show one ray for simplicity. We want to find how the focal length *FP* (denoted by *f*) relates to the radius of curvature of the mirror, *R*, whose length is $R = CF + FP$. The law of reflection tells us that angles *OXC* and *CXF* are the same, and because the incident ray is parallel to the optical axis, angles *OXC* and *XCP* are also the same. Thus, triangle *CXF* is an isosceles triangle with $CF = FX$. If the angle θ is small (so that $\sin\theta \approx \theta$; this is called the "small-angle approximation"), then $FX \approx FP$ or $CF \approx FP$. Inserting this into the equation for the radius *R*, we get

Equation:

$$R = CF + FP = FP + FP = 2FP = 2f$$



Reflection in a concave mirror. In the small-angle approximation, a ray that is parallel to the optical axis *CP* is reflected through the focal point *F* of the mirror.

In other words, in the small-angle approximation, the focal length *f* of a concave spherical mirror is half of its radius of curvature, *R*:

Note:

Equation:

$$f = \frac{R}{2}.$$

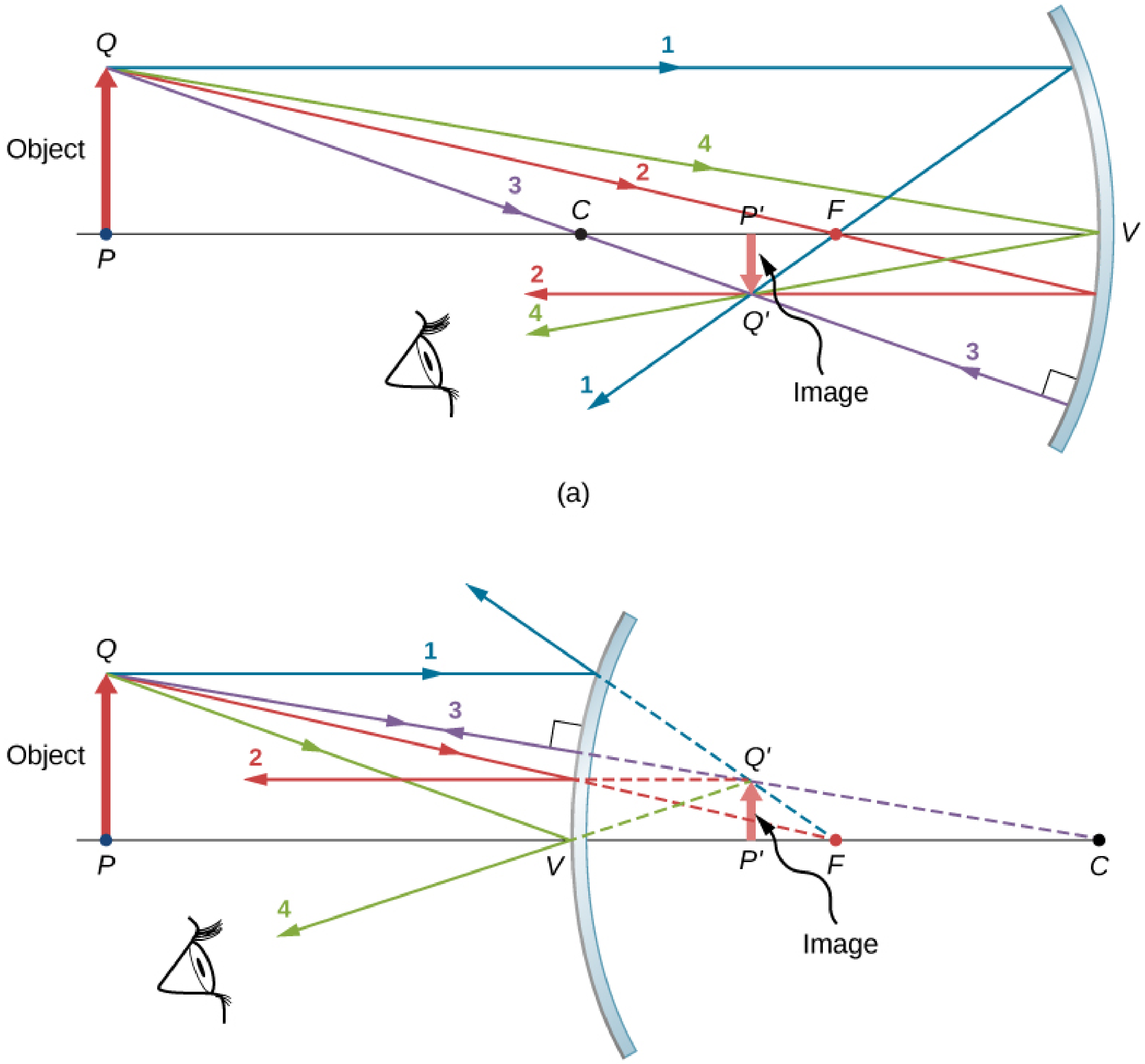
In this chapter, we assume that the **small-angle approximation** (also called the paraxial approximation) is always valid. In this approximation, all rays are paraxial rays, which means that they make a small angle with the optical axis and are at a distance much less than the radius of curvature from the optical axis. In this case, their angles θ of reflection are small angles, so $\sin \theta \approx \tan \theta \approx \theta$.

Using Ray Tracing to Locate Images

To find the location of an image formed by a spherical mirror, we first use ray tracing, which is the technique of drawing rays and using the law of reflection to determine the reflected rays (later, for lenses, we use the law of refraction to determine refracted rays). Combined with some basic geometry, we can use ray tracing to find the focal point, the image location, and other information about how a mirror manipulates light. In fact, we already used ray tracing above to locate the focal point of spherical mirrors, or the image distance of flat mirrors. To locate the image of an object, you must locate at least two points of the image. Locating each point requires drawing at least two rays from a point on the object and constructing their reflected rays. The point at which the reflected rays intersect, either in real space or in virtual space, is where the corresponding point of the image is located. To make ray tracing easier, we concentrate on four “principal” rays whose reflections are easy to construct.

[link] shows a concave mirror and a convex mirror, each with an arrow-shaped object in front of it. These are the objects whose images we want to locate by ray tracing. To do so, we draw rays from point Q that is on the object but not on the optical axis. We choose to draw our ray from the tip of the object. Principal ray 1 goes from point Q and travels parallel to the optical axis. The reflection of this ray must pass through the focal point, as discussed above. Thus, for the concave mirror, the reflection of principal

ray 1 goes through focal point *F*, as shown in part (b) of the figure. For the convex mirror, the backward extension of the reflection of principal ray 1 goes through the focal point (i.e., a virtual focus). Principal ray 2 travels first on the line going through the focal point and then is reflected back along a line parallel to the optical axis. Principal ray 3 travels toward the center of curvature of the mirror, so it strikes the mirror at normal incidence and is reflected back along the line from which it came. Finally, principal ray 4 strikes the vertex of the mirror and is reflected symmetrically about the optical axis.


Q
1
Object
4
2
3
C
P'
F
V
P
2
4
Q'
1
Image
3
Q
1
3
Object
2
Q'
P
V
P'
F
C
4
Image

(a)

(b)

The four principal rays shown for both (a) a concave mirror and (b) a convex mirror. The image forms where the rays intersect (for real images) or where their backward extensions intersect (for virtual images).

The four principal rays intersect at point Q', which is where the image of point Q is located. To locate point Q', drawing any two of these principle rays would suffice. We are thus free to choose whichever of the principal rays we desire to locate the image. Drawing more than two principal rays is sometimes useful to verify that the ray tracing is correct.

To completely locate the extended image, we need to locate a second point in the image, so that we know how the image is oriented. To do this, we trace the principal rays from the base of the object. In this case, all four principal rays run along the optical axis, reflect from the mirror, and then run back along the optical axis. The difficulty is that, because these rays are collinear, we cannot determine a unique point where they intersect. All we know is that the base of the image is on the optical axis. However, because the mirror is symmetrical from top to bottom, it does not change the vertical orientation of the object. Thus, because the object is vertical, the image must be vertical. Therefore, the image of the base of the object is on the optical axis directly above the image of the tip, as drawn in the figure.

For the concave mirror, the extended image in this case forms between the focal point and the center of curvature of the mirror. It is inverted with respect to the object, is a real image, and is smaller than the object. Were we to move the object closer to or farther from the mirror, the characteristics of the image would change. For example, we show, as a later exercise, that an object placed between a concave mirror and its focal point leads to a virtual image that is upright and larger than the object. For the convex mirror, the extended image forms between the focal point and the mirror. It is upright with respect to the object, is a virtual image, and is smaller than the object.

Summary of Ray-Tracing Rules

Ray tracing is very useful for mirrors. The rules for ray tracing are summarized here for reference:

- A ray travelling parallel to the optical axis of a spherical mirror is reflected along a line that goes through the focal point of the mirror (ray 1 in [link]).
- A ray travelling along a line that goes through the focal point of a spherical mirror is reflected along a line parallel to the optical axis of the mirror (ray 2 in [link]).

- A ray travelling along a line that goes through the center of curvature of a spherical mirror is reflected back along the same line (ray 3 in [link]).
- A ray that strikes the vertex of a spherical mirror is reflected symmetrically about the optical axis of the mirror (ray 4 in [link]).

We use ray tracing to illustrate how images are formed by mirrors and to obtain numerical information about optical properties of the mirror. If we assume that a mirror is small compared with its radius of curvature, we can also use algebra and geometry to derive a mirror equation, which we do in the next section. Combining ray tracing with the mirror equation is a good way to analyze mirror systems.

Image Formation by Reflection—The Mirror Equation

For a plane mirror, we showed that the image formed has the same height and orientation as the object, and it is located at the same distance behind the mirror as the object is in front of the mirror. Although the situation is a bit more complicated for curved mirrors, using geometry leads to simple formulas relating the object and image distances to the focal lengths of concave and convex mirrors.

Consider the object *OP* shown in [link]. The center of curvature of the mirror is labeled *C* and is a distance *R* from the vertex of the mirror, as marked in the figure. The object and image distances are labeled d_{o} and d_{i}, and the object and image heights are labeled h_{o} and h_{i}, respectively. Because the angles ϕ and ϕ' are alternate interior angles, we know that they have the same magnitude. However, they must differ in sign if we measure angles from the optical axis, so $\phi = -\phi'$. An analogous scenario holds for the angles θ and θ'. The law of reflection tells us that they have the same magnitude, but their signs must differ if we measure angles from the optical axis. Thus, $\theta = -\theta'$. Taking the tangent of the angles θ and θ', and using the property that $\tan(-\theta) = -\tan\theta$, gives us

Equation:

$$\left.\begin{array}{l}\tan\theta = \frac{h_\text{o}}{d_\text{o}}\\ \tan\theta' = -\tan\theta = \frac{h_\text{i}}{d_\text{i}}\end{array}\right\}\frac{h_\text{o}}{d_\text{o}} = -\frac{h_\text{i}}{d_\text{i}}\text{ or } -\frac{h_\text{o}}{h_\text{i}} = \frac{d_\text{o}}{d_\text{i}}.$$

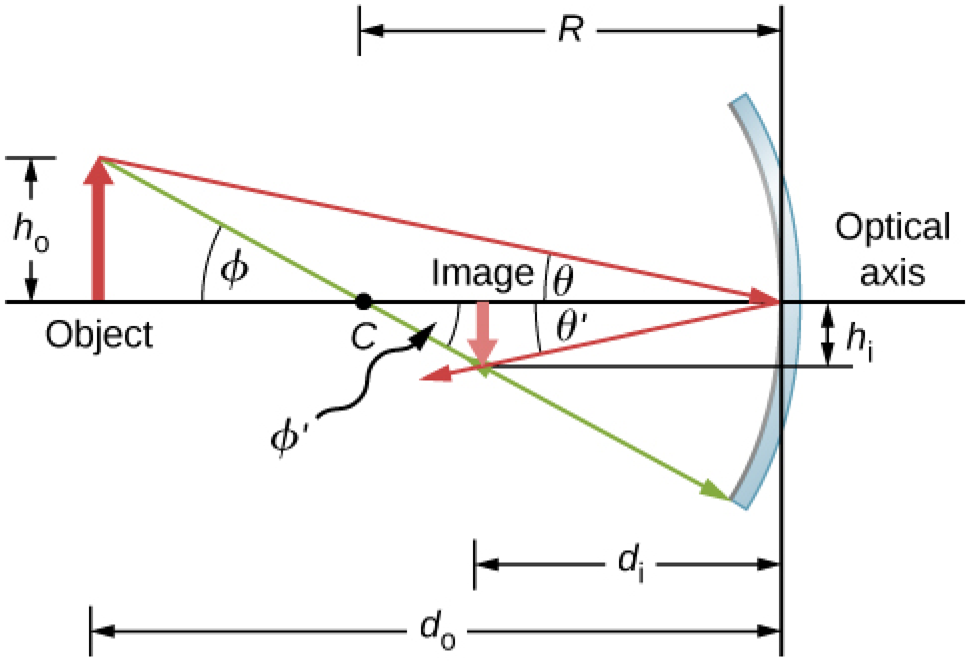

R
h_o
Object
ϕ
C
Image
θ
θ'
Optical axis
h_i
ϕ'
d_i
d_o

Image formed by a concave mirror.

Similarly, taking the tangent of ϕ and ϕ' gives
Equation:

$$\left.\begin{array}{l}\tan\phi = \frac{h_\text{o}}{d_\text{o}-R}\\ \tan\phi' = -\tan\phi = \frac{h_\text{i}}{R-d_\text{i}}\end{array}\right\}\frac{h_\text{o}}{d_\text{o}-R} = -\frac{h_\text{i}}{R-d_\text{i}}\text{ or } -\frac{h_\text{o}}{h_\text{i}} = \frac{d_\text{o}-R}{R-d_\text{i}}.$$

Combining these two results gives
Equation:

$$\frac{d_{\text{o}}}{d_{\text{i}}} = \frac{d_{\text{o}} - R}{R - d_{\text{i}}}.$$

After a little algebra, this becomes
Equation:

$$\frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{2}{R}.$$

No approximation is required for this result, so it is exact. However, as discussed above, in the small-angle approximation, the focal length of a spherical mirror is one-half the radius of curvature of the mirror, or $f = R/2$. Inserting this into [link] gives the *mirror equation*:

Note:
Equation:

$$\frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{f}.$$

The mirror equation relates the image and object distances to the focal distance and is valid only in the small-angle approximation. Although it was derived for a concave mirror, it also holds for convex mirrors (proving this is left as an exercise). We can extend the mirror equation to the case of a plane mirror by noting that a plane mirror has an infinite radius of curvature. This means the focal point is at infinity, so the mirror equation simplifies to
Equation:

$$d_{\text{o}} = -d_{\text{i}}$$

which is the same as [link] obtained earlier.

Notice that we have been very careful with the signs in deriving the mirror equation. For a plane mirror, the image distance has the opposite sign of the object distance. Also, the real image formed by the concave mirror in [link] is on the opposite side of the optical axis with respect to the object. In this case, the image height should have the opposite sign of the object height. To keep track of the signs of the various quantities in the mirror equation, we now introduce a sign convention.

Sign convention for spherical mirrors

Using a consistent sign convention is very important in geometric optics. It assigns positive or negative values for the quantities that characterize an optical system. Understanding the sign convention allows you to describe an image without constructing a ray diagram. This text uses the following sign convention:

1. The focal length f is positive for concave mirrors and negative for convex mirrors.
2. The image distance d_i is positive for real images and negative for virtual images.

Notice that rule 1 means that the radius of curvature of a spherical mirror can be positive or negative. What does it mean to have a negative radius of curvature? This means simply that the radius of curvature for a convex mirror is defined to be negative.

Image magnification

Let’s use the sign convention to further interpret the derivation of the mirror equation. In deriving this equation, we found that the object and image heights are related by

Equation:

$$-\frac{h_\text{o}}{h_\text{i}} = \frac{d_\text{o}}{d_\text{i}}.$$

See [link]. Both the object and the image formed by the mirror in [link] are real, so the object and image distances are both positive. The highest point of the object is above the optical axis, so the object height is positive. The image, however, is below the optical axis, so the image height is negative. Thus, this sign convention is consistent with our derivation of the mirror equation.

[link] in fact describes the **linear magnification** (often simply called “magnification”) of the image in terms of the object and image distances. We thus define the dimensionless magnification *m* as follows:
Equation:

$$m = \frac{h_\text{i}}{h_\text{o}}.$$

If *m* is positive, the image is upright, and if *m* is negative, the image is inverted. If $|m| > 1$, the image is larger than the object, and if $|m| < 1$, the image is smaller than the object. With this definition of magnification, we get the following relation between the vertical and horizontal object and image distances:

Note:
Equation:

$$m = \frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}}.$$

This is a very useful relation because it lets you obtain the magnification of the image from the object and image distances, which you can obtain from the mirror equation.

Example:

Solar Electric Generating System

One of the solar technologies used today for generating electricity involves a device (called a parabolic trough or concentrating collector) that concentrates sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where the thermal energy is transferred to another system that is used to generate steam and eventually generates electricity through a conventional steam cycle. [link] shows such a working system in southern California. The real mirror is a parabolic cylinder with its focus located at the pipe; however, we can approximate the mirror as exactly one-quarter of a circular cylinder.

Parabolic trough collectors are used to generate electricity in southern California. (credit: “kjkolb”/Wikimedia Commons)

a. If we want the rays from the sun to focus at 40.0 cm from the mirror, what is the radius of the mirror?
b. What is the amount of sunlight concentrated onto the pipe, per meter of pipe length, assuming the insolation (incident solar radiation) is $900\ \mathrm{W/m^2}$?

c. If the fluid-carrying pipe has a 2.00-cm diameter, what is the temperature increase of the fluid per meter of pipe over a period of 1 minute? Assume that all solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

First identify the physical principles involved. Part (a) is related to the optics of spherical mirrors. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution

a. The sun is the object, so the object distance is essentially infinity: $d_\text{o} = \infty$. The desired image distance is $d_\text{i} = 40.0\text{ cm}$. We use the mirror equation to find the focal length of the mirror:

   Equation:

$$\begin{aligned} \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} &= \frac{1}{f} \\ f &= \left(\frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}\right)^{-1} \\ &= \left(\frac{1}{\infty} + \frac{1}{40.0\text{ cm}}\right)^{-1} \\ &= 40.0\text{ cm} \end{aligned}$$

   Thus, the radius of the mirror is $R = 2f = 80.0\text{ cm}$.

b. The insolation is $900\ \text{W/m}^2$. You must find the cross-sectional area A of the concave mirror, since the power delivered is $900\ \text{W/m}^2 \times A$. The mirror in this case is estimated as a quarter-section of a cylinder, so the area for a length L of the mirror is $A = \frac{1}{4}(2\pi R)L$. The area for a length of 1.00 m is then

   Equation:

$$A = \frac{\pi}{2}R(1.00\text{ m}) = \frac{(3.14)}{2}(0.800\text{ m})(1.00\text{ m}) = 1.26\text{ m}^2.$$

   The insolation on the 1.00-m length of pipe is then

Equation:

$$\left(9.00\ \times\ 10^2\frac{\text{W}}{\text{m}^2}\right)\left(1.26\,\text{m}^2\right) = 1130\,\text{W}.$$

c. The increase in temperature is given by $Q = mc\Delta T$. The mass *m* of the mineral oil in the one-meter section of pipe is

Equation:

$$\begin{aligned} m &= \rho V = \rho\pi\left(\tfrac{d}{2}\right)^2(1.00\,\text{m}) \\ &= \left(8.00\ \times\ 10^2\,\text{kg/m}^3\right)(3.14)(0.0100\,\text{m})^2(1.00\,\text{m}) \\ &= 0.251\,\text{kg} \end{aligned}$$

Therefore, the increase in temperature in one minute is

Equation:

$$\begin{aligned} \Delta T &= Q/mc \\ &= \frac{(1130\,\text{W})(60.0\,\text{s})}{(0.251\,\text{kg})(1670\,\text{J·kg/}^\circ\text{C})} \\ &= 162^\circ\text{C} \end{aligned}$$

Significance

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400°C. We are considering only one meter of pipe here and ignoring heat losses along the pipe.

Example:

Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea of the eye, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12 cm from the cornea and the image magnification is 0.032, what is the radius of curvature of the cornea?

Strategy

If you find the focal length of the convex mirror formed by the cornea, then you know its radius of curvature (it’s twice the focal length). The object distance is $d_\text{o} = 12\,\text{cm}$ and the magnification is $m = 0.032$. First find the image distance d_i and then solve for the focal length f.

Solution

Start with the equation for magnification, $m = -d_\text{i}/d_\text{o}$. Solving for d_i and inserting the given values yields

Equation:

$$d_\text{i} = -md_\text{o} = -(0.032)(12\,\text{cm}) = -0.384\,\text{cm}$$

where we retained an extra significant figure because this is an intermediate step in the calculation. Solve the mirror equation for the focal length f and insert the known values for the object and image distances. The result is

Equation:

$$\begin{aligned} \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} &= \frac{1}{f} \\ f &= \left(\frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}\right)^{-1} \\ &= \left(\frac{1}{12\,\text{cm}} + \frac{1}{-0.384\,\text{cm}}\right)^{-1} \\ &= -0.40\,\text{cm} \end{aligned}$$

The radius of curvature is twice the focal length, so

Equation:

$$R = 2f = -0.80\,\text{cm}$$

Significance

The focal length is negative, so the focus is virtual, as expected for a concave mirror and a real object. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, corneas may not be spherical, which complicates the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror. Thus, the image is virtual because no rays actually pass through it. In the problems and exercises, you will show that, for a fixed object distance, a smaller radius of curvature corresponds to a smaller the magnification.

Note:

Problem-Solving Strategy: Spherical Mirrors

Step 1. First make sure that image formation by a spherical mirror is involved.

Step 2. Determine whether ray tracing, the mirror equation, or both are required. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and known values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns).

Step 5. If ray tracing is required, use the ray-tracing rules listed near the beginning of this section.

Step 6. Most quantitative problems require using the mirror equation. Use the examples as guides for using the mirror equation.

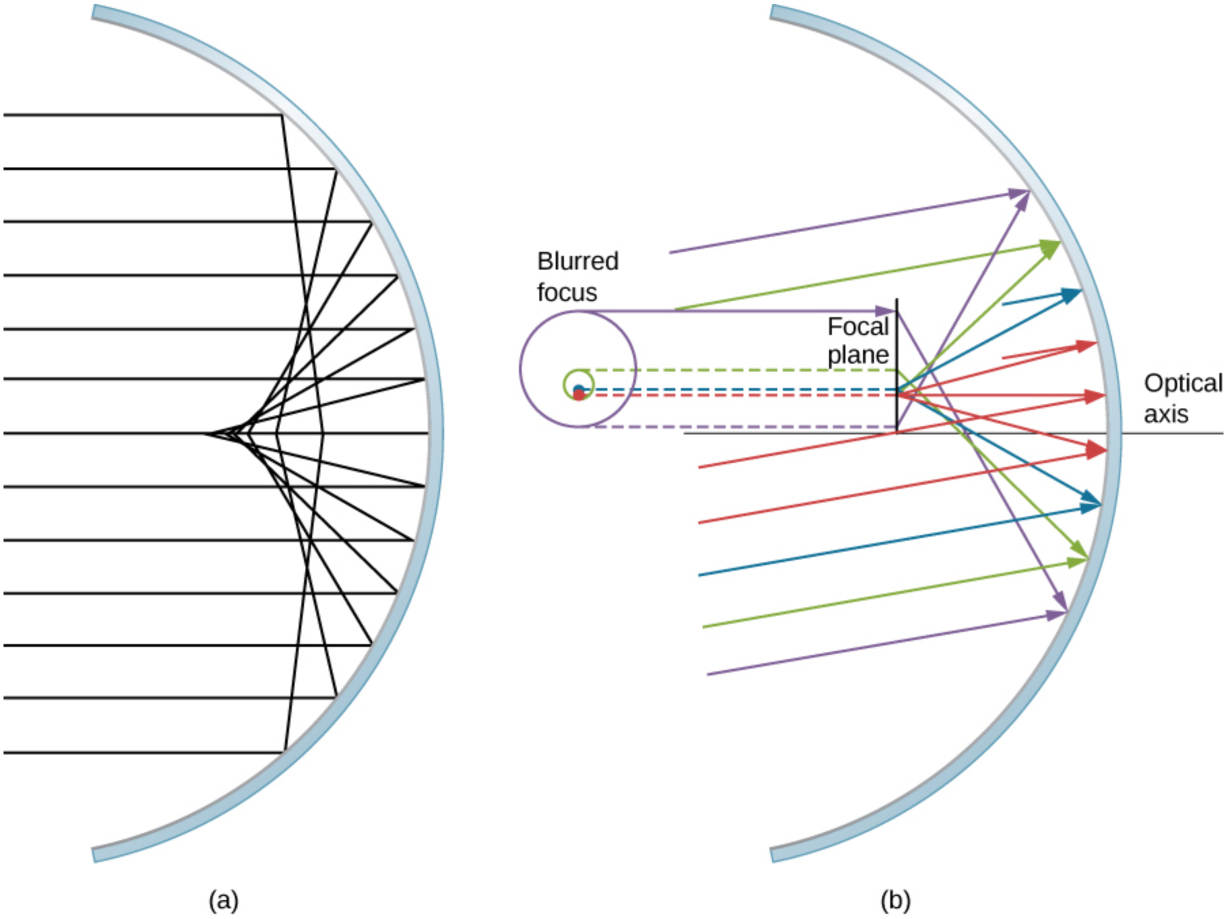
Step 7. Check to see whether the answer makes sense. Do the signs of object distance, image distance, and focal length correspond with what is expected from ray tracing? Is the sign of the magnification correct? Are the object and image distances reasonable?

Departure from the Small-Angle Approximation

The small-angle approximation is a cornerstone of the above discussion of image formation by a spherical mirror. When this approximation is violated, then the image created by a spherical mirror becomes distorted. Such distortion is called **aberration**. Here we briefly discuss two specific types of aberrations: spherical aberration and coma.

Spherical aberration

Consider a broad beam of parallel rays impinging on a spherical mirror, as shown in [link].



(a) With spherical aberration, the rays that are farther from the optical axis and the rays that are closer to the optical axis are focused at different points. Notice that the aberration gets worse for rays farther from the optical axis. (b) For comatic aberration, parallel rays that are not parallel to the optical axis are focused at different heights and at different focal lengths, so the image contains a “tail” like a comet (which is “coma” in Latin). Note that the colored rays are only to facilitate viewing; the colors do not indicate the color of the light.

The farther from the optical axis the rays strike, the worse the spherical mirror approximates a parabolic mirror. Thus, these rays are not focused at the same point as rays that are near the optical axis, as shown in the figure. Because of **spherical aberration**, the image of an extended object in a spherical mirror will be blurred. Spherical aberrations are characteristic of the mirrors and lenses that we consider in the following section of this chapter (more sophisticated mirrors and lenses are needed to eliminate spherical aberrations).

Coma or comatic aberration

Coma is similar to spherical aberration, but arises when the incoming rays are not parallel to the optical axis, as shown in part (b) of [link]. Recall that the small-angle approximation holds for spherical mirrors that are small compared to their radius. In this case, spherical mirrors are good approximations of parabolic mirrors. Parabolic mirrors focus all rays that are parallel to the optical axis at the focal point. However, parallel rays that are *not* parallel to the optical axis are focused at different heights and at different focal lengths, as show in part (b) of [link]. Because a spherical mirror is symmetric about the optical axis, the various colored rays in this figure create circles of the corresponding color on the focal plane.

Although a spherical mirror is shown in part (b) of [link], comatic aberration occurs also for parabolic mirrors—it does not result from a breakdown in the small-angle approximation. Spherical aberration, however, occurs only for spherical mirrors and is a result of a breakdown in the small-angle approximation. We will discuss both coma and spherical aberration later in this chapter, in connection with telescopes.

Summary

- Spherical mirrors may be concave (converging) or convex (diverging).
- The focal length of a spherical mirror is one-half of its radius of curvature: $f = R/2$.
- The mirror equation and ray tracing allow you to give a complete description of an image formed by a spherical mirror.
- Spherical aberration occurs for spherical mirrors but not parabolic mirrors; comatic aberration occurs for both types of mirrors.

Conceptual Questions

Exercise:

Problem: At what distance is an image always located: at d_o, d_i, or f?

Exercise:

Problem:

Under what circumstances will an image be located at the focal point of a spherical lens or mirror?

Solution:

when the object is at infinity; see the mirror equation

Exercise:

Problem:

What is meant by a negative magnification? What is meant by a magnification whose absolute value is less than one?

Exercise:

Problem:

Can an image be larger than the object even though its magnification is negative? Explain.

Solution:

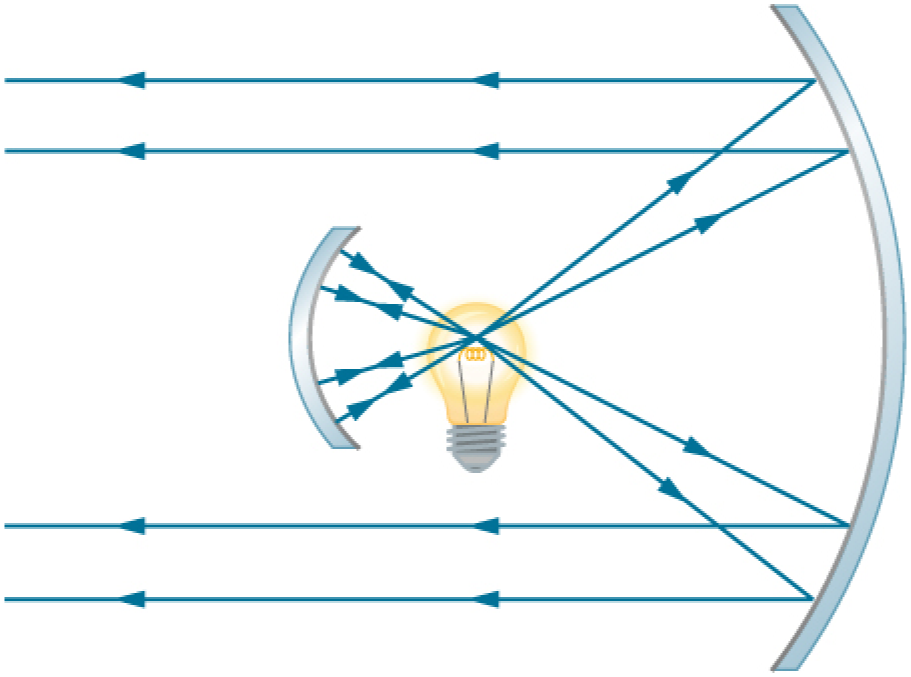
Yes, negative magnification simply means that the image is upside down; this does not prevent the image from being larger than the object. For instance, for a concave mirror, if distance to the object is larger than one focal distance but smaller than two focal distances the image will be inverted and magnified.

Problems

Exercise:

Problem:

The following figure shows a light bulb between two spherical mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?



Solution:

It is in the focal point of the big mirror and at the center of curvature of the small mirror.

Exercise:

Problem:

Why are diverging mirrors often used for rearview mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

Exercise:

Problem:

Some telephoto cameras use a mirror rather than a lens. What radius of curvature mirror is needed to replace a 800 mm-focal length telephoto lens?

Solution:

$$f=\frac{R}{2}\Rightarrow R=+1.60\,\mathrm{m}$$

Exercise:

Problem:

Calculate the focal length of a mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature.

Exercise:

Problem:

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR radiation follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Solution:

$d_\text{o} = 27.3\ \text{cm}$

Exercise:

Problem:

Find the magnification of the heater element in the previous problem. Note that its large magnitude helps spread out the reflected energy.

Exercise:

Problem:

What is the focal length of a makeup mirror that produces a magnification of 1.50 when a person’s face is 12.0 cm away? Explicitly show how you follow the steps in the [link].

Solution:

Step 1: Image formation by a mirror is involved.
Step 2: Draw the problem set up when possible.
Step 3: Use thin-lens equations to solve this problem.
Step 4: Find f.
Step 5: Given: $m = 1.50, d_\text{o} = 0.120\ \text{m}$.
Step 6: No ray tracing is needed.
Step 7: Using $m = \frac{d_\text{i}}{d_\text{o}}$, $d_\text{i} = -0.180\ \text{m}$. Then, $f = 0.360\ \text{m}$.
Step 8: The image is virtual because the image distance is negative. The focal length is positive, so the mirror is concave.

Exercise:

Problem:

A shopper standing 3.00 m from a convex security mirror sees his image with a magnification of 0.250. (a) Where is his image? (b) What is the focal length of the mirror? (c) What is its radius of curvature?

Exercise:

Problem:

An object 1.50 cm high is held 3.00 cm from a person's cornea, and its reflected image is measured to be 0.167 cm high. (a) What is the magnification? (b) Where is the image? (c) Find the radius of curvature of the convex mirror formed by the cornea. (Note that this technique is used by optometrists to measure the curvature of the cornea for contact lens fitting. The instrument used is called a keratometer, or curve measurer.)

Solution:

a. for a convex mirror $d_\text{i} < 0 \Rightarrow m > 0.$ $m = +0.111$; b. $d_\text{i} = -0.334\,\text{cm}$ (behind the cornea);
c. $f = -0.376\,\text{cm}$, so that $R = -0.752\,\text{cm}$

Exercise:

Problem:

Ray tracing for a flat mirror shows that the image is located a distance behind the mirror equal to the distance of the object from the mirror. This is stated as $d_\text{i} = -d_\text{o}$, since this is a negative image distance (it is a virtual image). What is the focal length of a flat mirror?

Exercise:

Problem:

Show that, for a flat mirror, $h_\text{i} = h_\text{o}$, given that the image is the same distance behind the mirror as the distance of the object from the mirror.

Solution:

$m = \frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = -\frac{-d_\text{o}}{d_\text{o}} = \frac{d_\text{o}}{d_\text{o}} = 1 \Rightarrow h_\text{i} = h_\text{o}$

Exercise:

Problem:

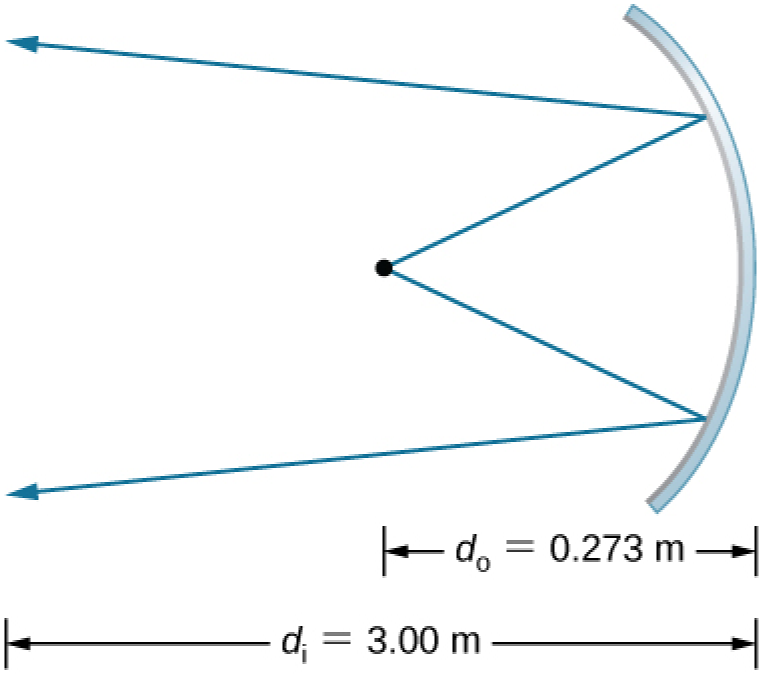
Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature. That is, prove that $f = R/2$. Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.

Exercise:

Problem:

Referring to the electric room heater considered in problem 5, calculate the intensity of IR radiation in $\mathrm{W/m^2}$ projected by the concave mirror on a person 3.00 m away. Assume that the heating element radiates 1500 W and has an area of $100\ \mathrm{cm}^2$, and that half of the radiated power is reflected and focused by the mirror.

Solution:



$m = -11.0$

$A' = 0.110\,\text{m}^2$

$I = 6.82\,\text{kW/m}^2$

Exercise:

Problem:

Two mirrors are inclined at an angle of 60° and an object is placed at a point that is equidistant from the two mirrors. Use a protractor to draw rays accurately and locate all images. You may have to draw several figures so that that rays for different images do not clutter your drawing.

Exercise:

Problem:

Two parallel mirrors are facing each other and are separated by a distance of 3 cm. A point object is placed between the mirrors 1 cm from one of the mirrors. Find the coordinates of all the images.

Solution:

$$x_{2m} = -x_{2m-1}, \quad (m = 1, 2, 3, \ldots),$$
$$x_{2m+1} = b - x_{2m}, \quad (m = 0, 1, 2, \ldots), \text{ with } x_0 = a.$$

Glossary

aberration	distortion in an image caused by departures from the small-angle approximation
coma	similar to spherical aberration, but arises when the incoming rays are not parallel to the optical axis

concave mirror

spherical mirror with its reflecting surface on the inner side of the sphere; the mirror forms a “cave”

convex mirror

spherical mirror with its reflecting surface on the outer side of the sphere

curved mirror

mirror formed by a curved surface, such as spherical, elliptical, or parabolic

focal length

distance along the optical axis from the focal point to the optical element that focuses the light rays

focal point

for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

linear magnification

ratio of image height to object height

optical axis

axis about which the mirror is rotationally symmetric; you can rotate the mirror about this axis without changing anything

small-angle approximation

approximation that is valid when the size of a spherical mirror is significantly smaller than the mirror’s radius; in this approximation, spherical aberration is negligible and the mirror has a well-defined focal point

spherical aberration

distortion in the image formed by a spherical mirror when rays are not all focused at the same point vertex

point where the mirror’s surface intersects with the optical axis

Images Formed by Refraction

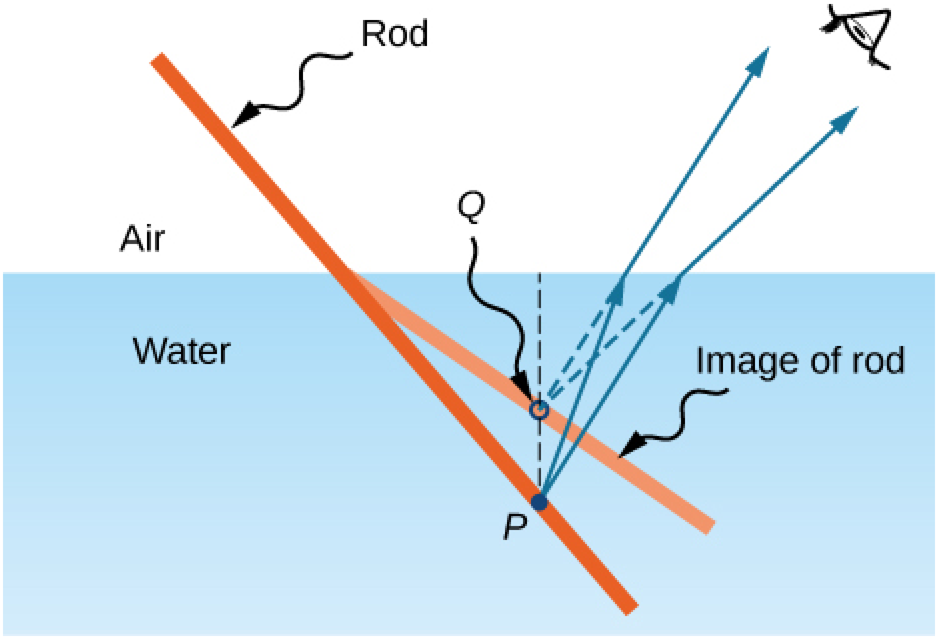
By the end of this section, you will be able to:

- Describe image formation by a single refracting surface
- Determine the location of an image and calculate its properties by using a ray diagram
- Determine the location of an image and calculate its properties by using the equation for a single refracting surface

When rays of light propagate from one medium to another, these rays undergo refraction, which is when light waves are bent at the interface between two media. The refracting surface can form an image in a similar fashion to a reflecting surface, except that the law of refraction (Snell’s law) is at the heart of the process instead of the law of reflection.

Refraction at a Plane Interface—Apparent Depth

If you look at a straight rod partially submerged in water, it appears to bend at the surface ([link]). The reason behind this curious effect is that the image of the rod inside the water forms a little closer to the surface than the actual position of the rod, so it does not line up with the part of the rod that is above the water. The same phenomenon explains why a fish in water appears to be closer to the surface than it actually is.

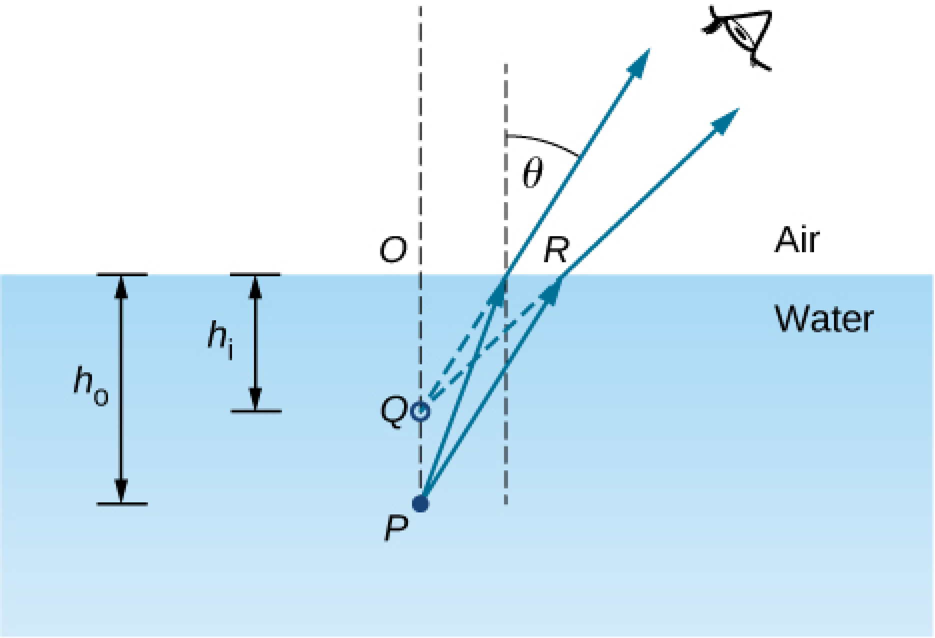


Bending of a rod at a water-air interface. Point *P* on the rod appears to be at point *Q*, which is where the image of point *P* forms due to refraction at the air-water interface.

To study image formation as a result of refraction, consider the following questions:

1. What happens to the rays of light when they enter or pass through a different medium?
2. Do the refracted rays originating from a single point meet at some point or diverge away from each other?

To be concrete, we consider a simple system consisting of two media separated by a plane interface ([link]). The object is in one medium and the observer is in the other. For instance, when you look at a fish from above the water surface, the fish is in medium 1 (the water) with refractive index 1.33, and your eye is in medium 2 (the air) with refractive index 1.00, and the surface of the water is the interface. The depth that you “see” is the image height h_{i} and is called the **apparent depth**. The actual depth of the fish is the object height h_{o}.



Apparent depth due to refraction. The real object at point P creates an image at point Q. The image is not at the same depth as the object, so the observer sees the image at an "apparent depth."

The apparent depth h_{i} depends on the angle at which you view the image. For a view from above (the so-called "normal" view), we can approximate the refraction angle θ to be small, and replace sin θ in Snell's law by tan θ. With this approximation, you can use the triangles ΔOPR and ΔOQR to show that the apparent depth is given by

Note:
Equation:

$$h_{\mathrm{i}} = \left(\frac{n_2}{n_1}\right) h_{\mathrm{o}}.$$

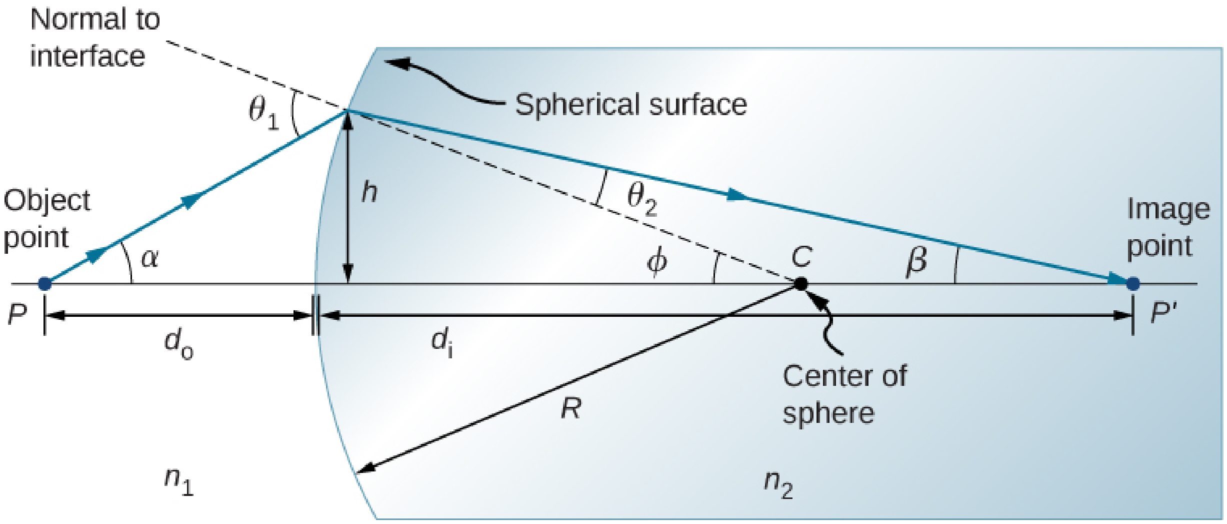
The derivation of this result is left as an exercise. Thus, a fish appears at 3/4 of the real depth when viewed from above.

Refraction at a Spherical Interface

Spherical shapes play an important role in optics primarily because high-quality spherical shapes are far easier to manufacture than other curved surfaces. To study refraction at a single spherical surface, we assume that the medium with the spherical surface at one end continues indefinitely (a “semi-infinite” medium).

Refraction at a convex surface

Consider a point source of light at point *P* in front of a convex surface made of glass (see [link]). Let *R* be the radius of curvature, n_1 be the refractive index of the medium in which object point *P* is located, and n_2 be the refractive index of the medium with the spherical surface. We want to know what happens as a result of refraction at this interface.


Normal to interface
θ_1
Spherical surface
Object point
α
h
θ_2
ϕ
C
β
Image point
P
P'
d_o
d_i
Center of sphere
R
n_1
n_2

Refraction at a convex surface $(n_2 > n_1)$.

Because of the symmetry involved, it is sufficient to examine rays in only one plane. The figure shows a ray of light that starts at the object point P, refracts at the interface, and goes through the image point P'. We derive a formula relating the object distance d_o, the image distance d_i, and the radius of curvature R.

Applying Snell's law to the ray emanating from point P gives $n_1 \sin\theta_1 = n_2 \sin\theta_2$. We work in the small-angle approximation, so $\sin\theta \approx \theta$ and Snell's law then takes the form

Equation:

$$n_1\theta_1 \approx n_2\theta_2.$$

From the geometry of the figure, we see that

Equation:

$$\theta_1 = \alpha + \phi, \quad \theta_2 = \phi - \beta.$$

Inserting these expressions into Snell's law gives

Equation:

$$n_1(\alpha + \phi) \approx n_2(\phi - \beta).$$

Using the diagram, we calculate the tangent of the angles $\alpha, \beta,$ and ϕ:

Equation:

$$\tan \alpha \approx \frac{h}{d_{\mathrm{o}}}, \quad \tan \beta \approx \frac{h}{d_{\mathrm{i}}}, \quad \tan \phi \approx \frac{h}{R}.$$

Again using the small-angle approximation, we find that $\tan \theta \approx \theta$, so the above relationships become

Equation:

$$\alpha \approx \frac{h}{d_{\mathrm{o}}}, \quad \beta \approx \frac{h}{d_{\mathrm{i}}}, \quad \phi \approx \frac{h}{R}.$$

Putting these angles into Snell’s law gives

Equation:

$$n_1 \left(\frac{h}{d_{\mathrm{o}}} + \frac{h}{R} \right) = n_2 \left(\frac{h}{R} - \frac{h}{d_{\mathrm{i}}} \right).$$

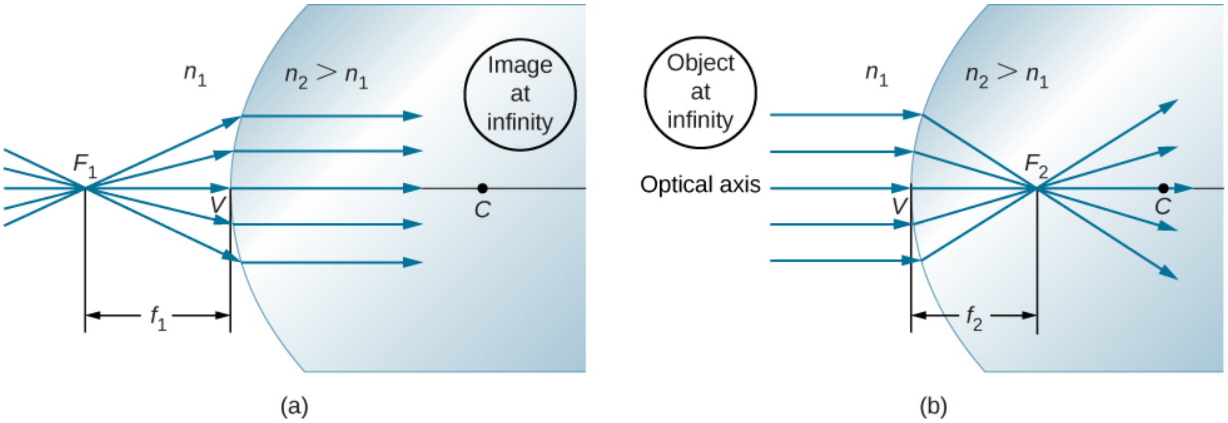
We can write this more conveniently as

Note:

Equation:

$$\frac{n_1}{d_{\mathrm{o}}} + \frac{n_2}{d_{\mathrm{i}}} = \frac{n_2 - n_1}{R}.$$

If the object is placed at a special point called the **first focus**, or the **object focus** F_1, then the image is formed at infinity, as shown in part (a) of [link].



n_1
$n_2 > n_1$
Image at infinity
F_1
V
C
f_1
(a)
Object at infinity
Optical axis
n_1
$n_2 > n_1$
F_2
V
C
f_2
(b)

(a) First focus (called the “object focus”) for refraction at a convex surface. (b) Second focus (called “image focus”) for refraction at a convex surface.

We can find the location f_1 of the first focus F_1 by setting $d_\mathrm{i} = \infty$ in the preceding equation.

Equation:

$$\frac{n_1}{f_1} + \frac{n_2}{\infty} = \frac{n_2 - n_1}{R}$$

Equation:

$$f_1 = \frac{n_1 R}{n_2 - n_1}$$

Similarly, we can define a **second focus** or **image focus** F_2 where the image is formed for an object that is far away [part (b)]. The location of the second focus F_2 is obtained from [link] by setting $d_\mathrm{o} = \infty$:

Equation:

$$\frac{n_1}{\infty} + \frac{n_2}{f_2} = \frac{n_2 - n_1}{R}$$

Equation:

$$f_2 = \frac{n_2 R}{n_2 - n_1}.$$

Note that the object focus is at a different distance from the vertex than the image focus because $n_1 \neq n_2$.

Sign convention for single refracting surfaces

Although we derived this equation for refraction at a convex surface, the same expression holds for a concave surface, provided we use the following sign convention:

1. $R > 0$ if surface is convex toward object; otherwise, $R < 0$.
2. $d_\mathrm{i} > 0$ if image is real and on opposite side from the object; otherwise, $d_\mathrm{i} < 0$.

Summary

This section explains how a single refracting interface forms images.

- When an object is observed through a plane interface between two media, then it appears at an apparent distance h_i that differs from the actual distance h_o: $h_\mathrm{i} = (n_2/n_1)h_\mathrm{o}$.
- An image is formed by the refraction of light at a spherical interface between two media of indices of refraction n_1 and n_2.
- Image distance depends on the radius of curvature of the interface, location of the object, and the indices of refraction of the media.

Conceptual Questions

Exercise:

Problem:

Derive the formula for the apparent depth of a fish in a fish tank using Snell’s law.

Exercise:

Problem:

Use a ruler and a protractor to find the image by refraction in the following cases. Assume an air-glass interface. Use a refractive index of 1 for air and of 1.5 for glass. (*Hint*: Use Snell’s law at the interface.)

(a) A point object located on the axis of a concave interface located at a point within the focal length from the vertex.

(b) A point object located on the axis of a concave interface located at a point farther than the focal length from the vertex.

(c) A point object located on the axis of a convex interface located at a point within the focal length from the vertex.

(d) A point object located on the axis of a convex interface located at a point farther than the focal length from the vertex.

(e) Repeat (a)–(d) for a point object off the axis.

Solution:

answers may vary

Problems

Exercise:

Problem:

An object is located in air 30 cm from the vertex of a concave surface made of glass with a radius of curvature 10 cm. Where does the image by refraction form and what is its magnification? Use $n_{\text{air}} = 1$ and $n_{\text{glass}} = 1.5$.

Exercise:

Problem:

An object is located in air 30 cm from the vertex of a convex surface made of glass with a radius of curvature 80 cm. Where does the image by refraction form and what is its magnification?

Solution:

$d_{\text{i}} = -55\text{ cm}; m = +1.8$

Exercise:

Problem:

An object is located in water 15 cm from the vertex of a concave surface made of glass with a radius of curvature 10 cm. Where does the image by refraction form and what is its magnification? Use $n_{\text{water}} = 4/3$ and $n_{\text{glass}} = 1.5$.

Exercise:

Problem:

An object is located in water 30 cm from the vertex of a convex surface made of Plexiglas with a radius of curvature of 80 cm. Where does the image form by refraction and what is its magnification? $n_{\text{water}} = 4/3$ and $n_{\text{Plexiglas}} = 1.65$.

Solution:

$d_{\text{i}} = -41\text{ cm}, m = 1.4$

Exercise:

Problem:

An object is located in air 5 cm from the vertex of a concave surface made of glass with a radius of curvature 20 cm. Where does the image form by refraction and what is its magnification? Use $n_{\text{air}} = 1$ and $n_{\text{glass}} = 1.5$.

Exercise:

Problem:

Derive the spherical interface equation for refraction at a concave surface. (*Hint*: Follow the derivation in the text for the convex surface.)

Solution:

proof

Glossary

apparent depth	depth at which an object is perceived to be located with respect to an interface between two media
first focus or object focus	object located at this point will result in an image created at infinity on the opposite side of a spherical interface between two media
second focus or image focus	for a converging interface, the point where a bundle of parallel rays refracting at a spherical interface; for a diverging interface, the point at which the backward continuation of the refracted rays will converge between two media will focus

The Eye

By the end of this section, you will be able to:

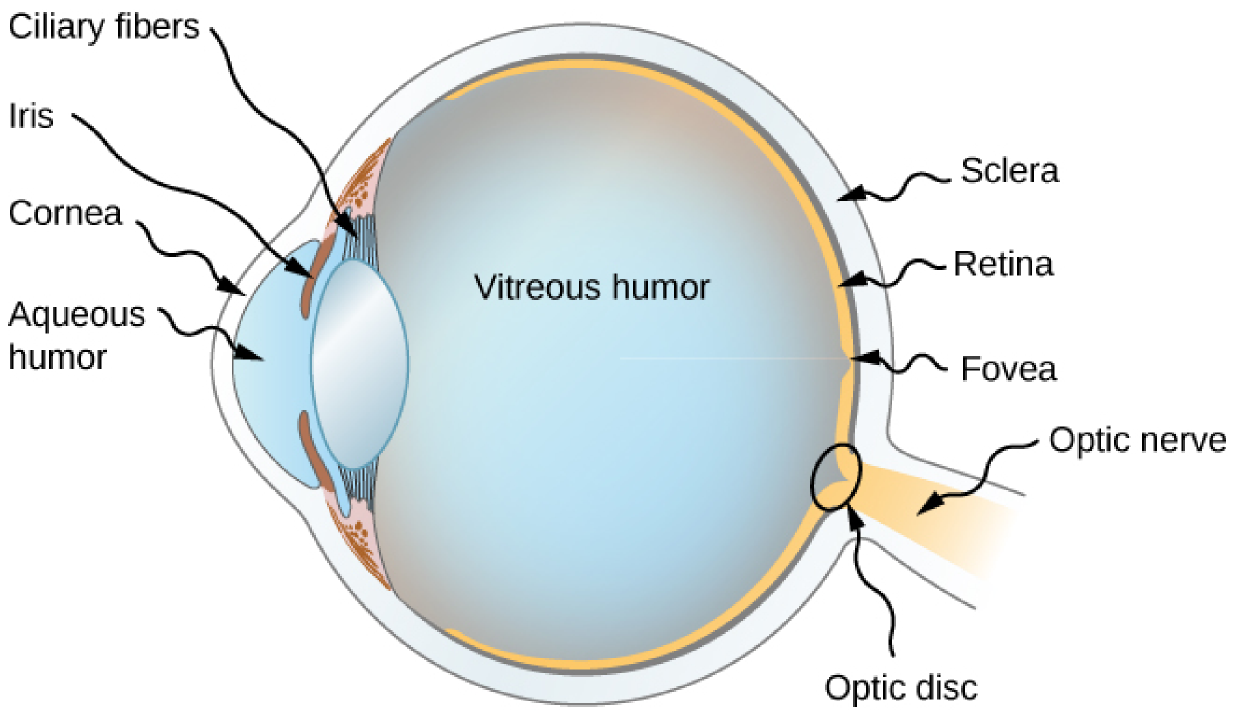
- Understand the basic physics of how images are formed by the human eye
- Recognize several conditions of impaired vision as well as the optics principles for treating these conditions

The human eye is perhaps the most interesting and important of all optical instruments. Our eyes perform a vast number of functions: They allow us to sense direction, movement, colors, and distance. In this section, we explore the geometric optics of the eye.

Physics of the Eye

The eye is remarkable in how it forms images and in the richness of detail and color it can detect. However, our eyes often need some correction to reach what is called “normal” vision. Actually, normal vision should be called “ideal” vision because nearly one-half of the human population requires some sort of eyesight correction, so requiring glasses is by no means “abnormal.” Image formation by our eyes and common vision correction can be analyzed with the optics discussed earlier in this chapter.

[link] shows the basic anatomy of the eye. The cornea and lens form a system that, to a good approximation, acts as a single thin lens. For clear vision, a real image must be projected onto the light-sensitive retina, which lies a fixed distance from the lens. The flexible lens of the eye allows it to adjust the radius of curvature of the lens to produce an image on the retina for objects at different distances. The center of the image falls on the fovea, which has the greatest density of light receptors and the greatest acuity (sharpness) in the visual field. The variable opening (i.e., the pupil) of the eye, along with chemical adaptation, allows the eye to detect light intensities from the lowest observable to 10^{10} times greater (without damage). This is an incredible range of detection. Processing of visual nerve impulses begins with interconnections in the retina and continues in the brain. The optic nerve conveys the signals received by the eye to the brain.

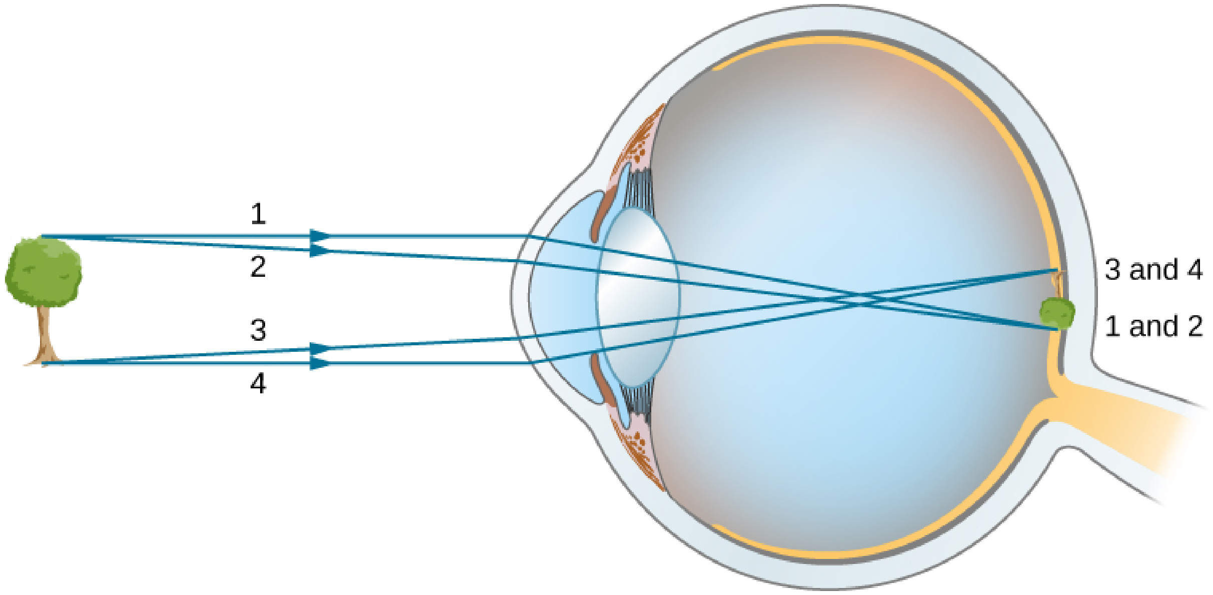


The cornea and lens of the eye act together to form a real image on the light-sensing retina, which has its densest concentration of receptors in the fovea and a blind spot over the optic nerve. The radius of curvature of the lens of an eye is adjustable to form an image on the retina for different object distances. Layers of tissues with varying indices of refraction in the lens are shown here. However, they have been omitted from other pictures for clarity.

The indices of refraction in the eye are crucial to its ability to form images. [link] lists the indices of refraction relevant to the eye. The biggest change in the index of refraction, which is where the light rays are most bent, occurs at the air-cornea interface rather than at the aqueous humor-lens interface. The ray diagram in [link] shows image formation by the cornea and lens of the eye. The cornea, which is itself a converging lens with a focal length of approximately 2.3 cm, provides most of the focusing power of the eye. The lens, which is a converging lens with a focal length of about 6.4 cm, provides the finer focus needed to produce a clear image on the retina. The cornea and lens can be treated as a single thin lens, even though the light rays pass through several layers of material (such as cornea, aqueous humor, several layers in the lens, and vitreous humor), changing direction at each interface. The image formed is much like the one produced by a single convex lens (i.e., a real, inverted image). Although images formed in the eye are inverted, the brain inverts them once more to make them seem upright.

Material	**Index of Refraction**
Water	1.33
Air	1.0
Cornea	1.38
Aqueous humor	1.34
Lens	1.41*
Vitreous humor	1.34

Refractive Indices Relevant to the Eye*This is an average value. The actual index of refraction varies throughout the lens and is greatest in center of the lens.



In the human eye, an image forms on the retina. Rays from the top and bottom of the object are traced to show how a real, inverted image is produced on the retina. The distance to the object is not to scale.

As noted, the image must fall precisely on the retina to produce clear vision—that is, the image distance d_i must equal the lens-to-retina distance. Because the lens-to-retina distance does not change, the image distance d_i must be the same for objects at all distances. The ciliary muscles adjust the shape of the eye lens for focusing on nearby or far objects. By changing the shape of the eye lens, the eye changes the focal length of the lens. This mechanism of the eye is called **accommodation**.

The nearest point an object can be placed so that the eye can form a clear image on the retina is called the **near point** of the eye. Similarly, the **far point** is the farthest distance at which an object is clearly visible. A person with normal vision can see objects clearly at distances ranging from 25 cm to essentially infinity. The near point increases with age, becoming several meters for some older people. In this text, we consider the near point to be 25 cm.

We can use the thin-lens equations to quantitatively examine image formation by the eye. First, we define the **optical power** of a lens as

Note:
Equation:

$$P = \frac{1}{f}$$

with the focal length *f* given in meters. The units of optical power are called “diopters” (D). That is, $1\,\text{D} = \frac{1}{\text{m}}$, or $1\,\text{m}^{-1}$. Optometrists prescribe common eyeglasses and contact lenses in units of diopters. With this definition of optical power, we can rewrite the thin-lens equations as

Equation:

$$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}.$$

Working with optical power is convenient because, for two or more lenses close together, the effective optical power of the lens system is approximately the sum of the optical power of the individual lenses:

Note:
Equation:

$$P_\text{total} = P_\text{lens 1} + P_\text{lens 2} + P_\text{lens 3} + \cdots$$

Example:

Effective Focal Length of the Eye

The cornea and eye lens have focal lengths of 2.3 and 6.4 cm, respectively. Find the net focal length and optical power of the eye.

Strategy

The optical powers of the closely spaced lenses add, so $P_{\text{eye}} = P_{\text{cornea}} + P_{\text{lens}}$.

Solution

Writing the equation for power in terms of the focal lengths gives

Equation:

$$\frac{1}{f_{\text{eye}}} = \frac{1}{f_{\text{cornea}}} + \frac{1}{f_{\text{lens}}} = \frac{1}{2.3\ \text{cm}} + \frac{1}{6.4\ \text{cm}}.$$

Hence, the focal length of the eye (cornea and lens together) is

Equation:

$$f_{\text{eye}} = 1.69\ \text{cm}.$$

The optical power of the eye is

Equation:

$$P_{\text{eye}} = \frac{1}{f_{\text{eye}}} = \frac{1}{0.0169\ \text{m}} = 59\ \text{D}.$$

For clear vision, the image distance d_i must equal the lens-to-retina distance. Normal vision is possible for objects at distances $d_\text{o} = 25\ \text{cm}$ to infinity. The following example shows how to calculate the image distance for an object placed at the near point of the eye.

Example:

Image of an object placed at the near point

The net focal length of a particular human eye is 1.7 cm. An object is placed at the near point of the eye. How far behind the lens is a focused image formed?

Strategy

The near point is 25 cm from the eye, so the object distance is $d_\text{o} = 25 \text{ cm}$. We determine the image distance from the lens equation:

Equation:

$$\frac{1}{d_\text{i}} = \frac{1}{f} - \frac{1}{d_\text{o}}.$$

Solution

Equation:

$$\begin{aligned} d_\text{i} &= \left(\frac{1}{f} - \frac{1}{d_\text{o}}\right)^{-1} \\ &= \left(\frac{1}{1.7 \text{ cm}} - \frac{1}{25 \text{ cm}}\right)^{-1} \\ &= 1.8 \text{ cm} \end{aligned}$$

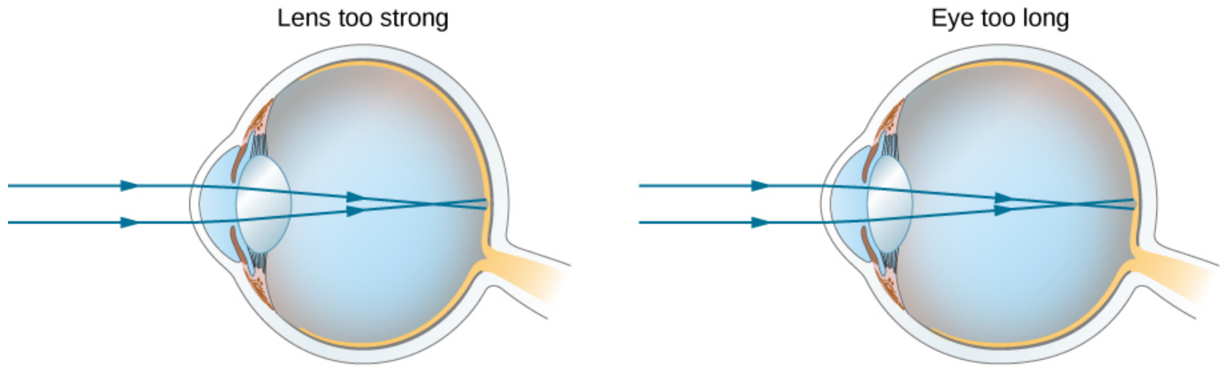
Therefore, the image is formed 1.8 cm behind the lens.

Significance

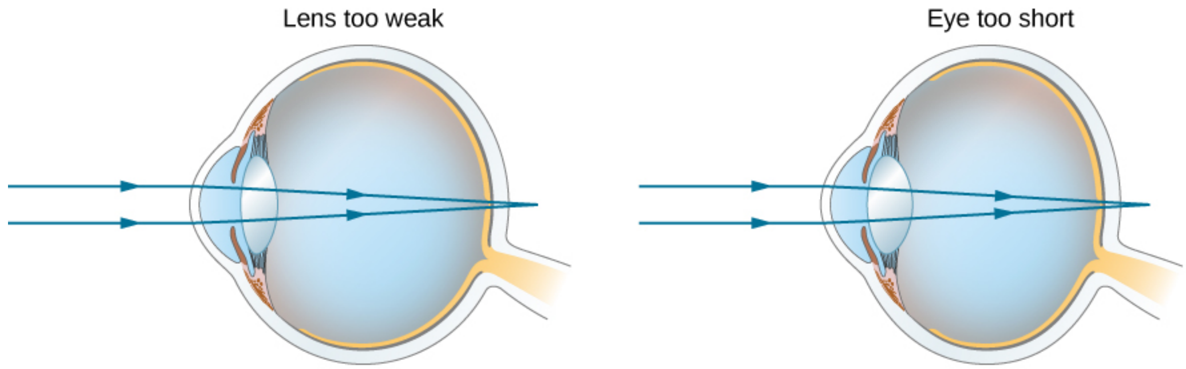
From the magnification formula, we find $m = -\frac{1.8 \text{ cm}}{25 \text{ cm}} = -0.073$. Since $m < 0$, the image is inverted in orientation with respect to the object. From the absolute value of *m* we see that the image is much smaller than the object; in fact, it is only 7% of the size of the object.

Vision Correction

The need for some type of vision correction is very common. Typical vision defects are easy to understand with geometric optics, and some are simple to correct. [link] illustrates two common vision defects. **Nearsightedness**, or **myopia**, is the ability to see near objects, whereas distant objects are blurry. The eye overconverges the nearly parallel rays from a distant object, and the rays cross in front of the retina. More divergent rays from a close object are converged on the retina for a clear image. The distance to the farthest object that can be seen clearly is called the far point of the eye (normally the far point is at infinity). **Farsightedness**, or **hyperopia**, is the ability to see far objects clearly, whereas near objects are blurry. A farsighted eye does not sufficiently converge the rays from a near object to make the rays meet on the retina.



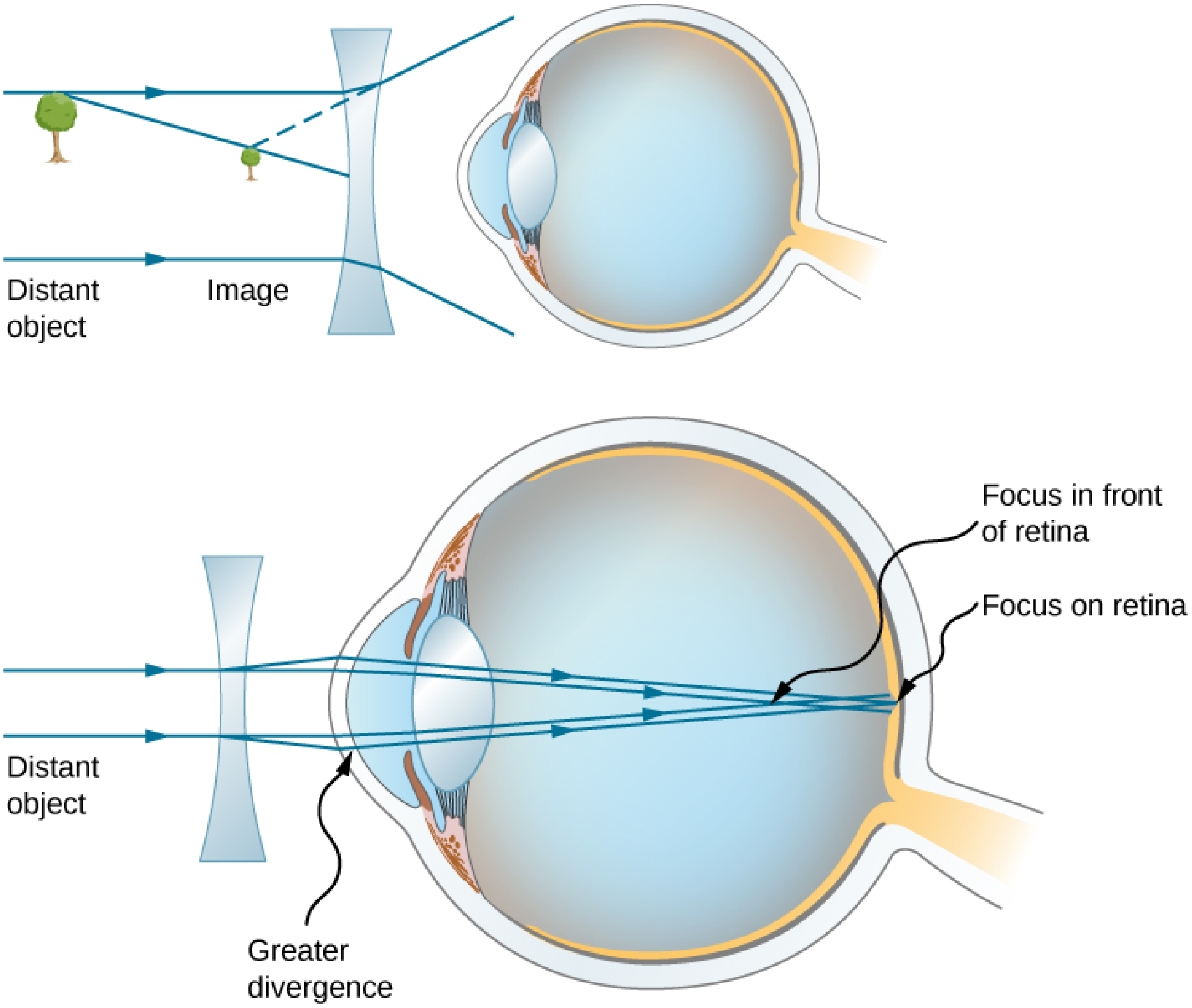
(a) Myopia



(b) Hyperopia

(a) The nearsighted (myopic) eye converges rays from a distant object in front of the retina, so they have diverged when they strike the retina, producing a blurry image. An eye lens that is too powerful can cause nearsightedness, or the eye may be too long. (b) The farsighted (hyperopic) eye is unable to converge the rays from a close object on the retina, producing blurry near-field vision. An eye lens with insufficient optical power or an eye that is too short can cause farsightedness.

Since the nearsighted eye overconverges light rays, the correction for nearsightedness consists of placing a diverging eyeglass lens in front of the eye, as shown in [link]. This reduces the optical power of an eye that is too powerful (recall that the focal length of a diverging lens is negative, so its optical power is negative). Another way to understand this correction is that a diverging lens will cause the incoming rays to diverge more to compensate for the excessive convergence caused by the lens system of the eye. The image produced by the diverging eyeglass lens serves as the (optical) object for the eye, and because the eye cannot focus on objects beyond its far point, the diverging lens must form an image of distant (physical) objects at a point that is closer than the far point.


Distant object
Image
Focus in front of retina
Focus on retina
Distant object
Greater divergence

Correction of nearsightedness requires a diverging lens that compensates for overconvergence by the eye. The diverging lens produces an image closer to the eye than the physical object. This image serves as the optical object for the eye, and the nearsighted person can see it clearly because it is closer than their far point.

Example:

Correcting Nearsightedness

What optical power of eyeglass lens is needed to correct the vision of a nearsighted person whose far point is 30.0 cm? Assume the corrective lens is fixed 1.50 cm away from the eye.

Strategy

You want this nearsighted person to be able to see distant objects clearly, which means that the eyeglass lens must produce an image 30.0 cm from the eye for an object at infinity. An image 30.0 cm from the eye will be $30.0\text{ cm} - 1.50\text{ cm} = 28.5\text{ cm}$ from the eyeglass lens. Therefore, we must have $d_\text{i} = -28.5\text{ cm}$ when $d_\text{o} = \infty$. The image distance is negative because it is on the same side of the eyeglass lens as the object.

Solution

Since d_i and d_o are known, we can find the optical power of the eyeglass lens by using [link]:

Equation:

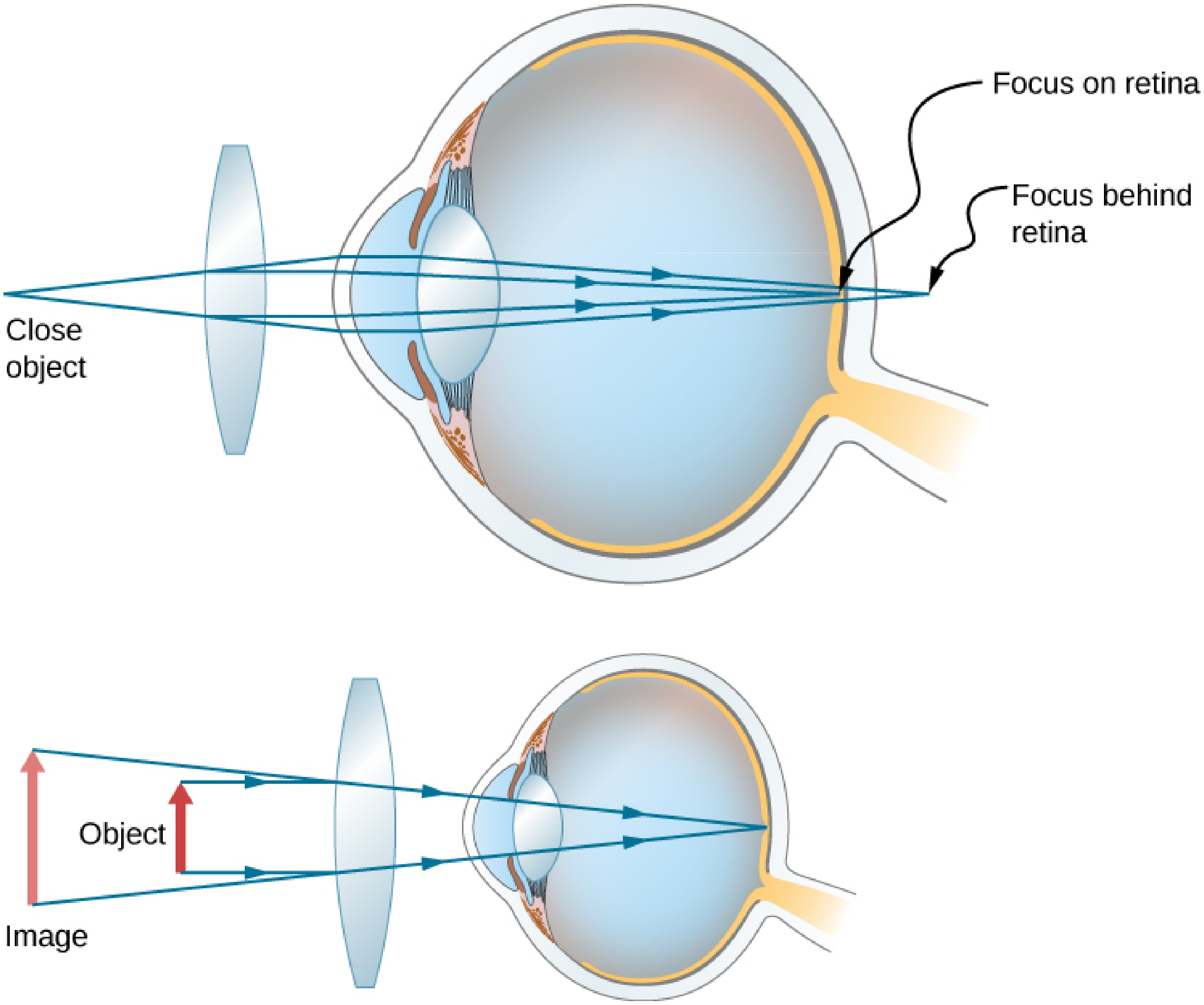
$$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} = \frac{1}{\infty} + \frac{1}{-0.285\text{ m}} = -3.51\text{D}.$$

Significance

The negative optical power indicates a diverging (or concave) lens, as expected. If you examine eyeglasses for nearsighted people, you will find the lenses are thinnest in the center. Additionally, if you examine a prescription for eyeglasses for nearsighted people, you will find that the prescribed optical power is negative and given in units of diopters.

Correcting farsightedness consists simply of using the opposite type of lens as for nearsightedness (i.e., a converging lens), as shown in [link].

Such a lens will produce an image of physical objects that are closer than the near point at a distance that is between the near point and the far point, so that the person can see the image clearly. To determine the optical power needed for correction, you must therefore know the person’s near point, as explained in [link].



Correction of farsightedness uses a converging lens that compensates for the underconvergence by the eye. The converging lens produces an image farther from the eye than the object, so that the farsighted person can see it clearly.

Example:

Correcting Farsightedness

What optical power of eyeglass lens is needed to allow a farsighted person, whose near point is 1.00 m, to see an object clearly that is 25.0 cm from the eye? Assume the corrective lens is fixed 1.5 cm from the eye.

Strategy

When an object is 25.0 cm from the person's eyes, the eyeglass lens must produce an image 1.00 m away (the near point), so that the person can see it clearly. An image 1.00 m from the eye will be $100\text{ cm}-1.5\text{ cm}=98.5\text{ cm}$ from the eyeglass lens because the eyeglass lens is 1.5 cm from the eye. Therefore, $d_\text{i}=-98.5\text{ cm}$, where the minus sign indicates that the image is on the same side of the lens as the object. The object is $25.0\text{ cm}-1.5\text{ cm}=23.5\text{ cm}$ from the eyeglass lens, so $d_\text{o}=23.5\text{ cm}$.

Solution

Since d_i and d_o are known, we can find the optical power of the eyeglass lens by using [link]:

Equation:

$$P=\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}=\frac{1}{0.235\text{ m}}+\frac{1}{-0.985\text{ m}}=+3.24\,\text{D}.$$

Significance

The positive optical power indicates a converging (convex) lens, as expected. If you examine eyeglasses of farsighted people, you will find the lenses to be thickest in the center. In addition, prescription eyeglasses for farsighted people have a prescribed optical power that is positive.

Summary

- Image formation by the eye is adequately described by the thin-lens equation.
- The eye produces a real image on the retina by adjusting its focal length in a process called accommodation.
- Nearsightedness, or myopia, is the inability to see far objects and is corrected with a diverging lens to reduce the optical power of the eye.
- Farsightedness, or hyperopia, is the inability to see near objects and is corrected with a converging lens to increase the optical power of the eye.
- In myopia and hyperopia, the corrective lenses produce images at distances that fall between the person's near and far points so that images can be seen clearly.

Conceptual Questions

Exercise:

Problem:

If the lens of a person's eye is removed because of cataracts (as has been done since ancient times), why would you expect an eyeglass lens of about 16 D to be prescribed?

Exercise:

Problem:

When laser light is shone into a relaxed normal-vision eye to repair a tear by spot-welding the retina to the back of the eye, the rays entering the eye must be parallel. Why?

Solution:

A relaxed, normal-vision eye will focus parallel rays of light onto the retina.

Exercise:

Problem:

Why is your vision so blurry when you open your eyes while swimming under water? How does a face mask enable clear vision?

Exercise:

Problem:

It has become common to replace the cataract-clouded lens of the eye with an internal lens. This intraocular lens can be chosen so that the person has perfect distant vision. Will the person be able to read without glasses? If the person was nearsighted, is the power of the intraocular lens greater or less than the removed lens?

Solution:

A person with an internal lens will need glasses to read because their muscles cannot distort the lens as they do with biological lenses, so they cannot focus on near objects. To correct nearsightedness, the power of the intraocular lens must be less than that of the removed lens.

Exercise:

Problem:

If the cornea is to be reshaped (this can be done surgically or with contact lenses) to correct myopia, should its curvature be made greater or smaller? Explain.

Problems

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

Exercise:

Problem:

What is the power of the eye when viewing an object 50.0 cm away?

Solution:

$$P = 52.0\,\text{D}$$

Exercise:

Problem:

Calculate the power of the eye when viewing an object 3.00 m away.

Exercise:

Problem:

The print in many books averages 3.50 mm in height. How high is the image of the print on the retina when the book is held 30.0 cm from the eye?

Solution:

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} \Rightarrow h_\text{i} = -h_\text{o}\left(\frac{d_\text{i}}{d_\text{o}}\right) = -(3.50\,\text{mm})\left(\frac{2.00\,\text{cm}}{30.0\,\text{cm}}\right) = -0.233\,\text{mm}$$

Exercise:

Problem:

Suppose a certain person's visual acuity is such that he can see objects clearly that form an image 4.00 µm high on his retina. What is the maximum distance at which he can read the 75.0-cm-high letters on the side of an airplane?

Exercise:

Problem:

People who do very detailed work close up, such as jewelers, often can see objects clearly at much closer distance than the normal 25 cm. (a) What is the power of the eyes of a woman who can see an object clearly at a distance of only 8.00 cm? (b) What is the image size of a 1.00-mm object, such as lettering inside a ring, held at this distance? (c) What would the size of the image be if the object were held at the normal 25.0 cm distance?

Solution:

a. $P = +62.5\text{ D}$;
b. $\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} \Rightarrow h_\text{i} = -0.250\,\text{mm}$;
c. $h_\text{i} = -0.0800\,\text{mm}$

Exercise:

Problem:

What is the far point of a person whose eyes have a relaxed power of 50.5 D?

Exercise:

Problem:

What is the near point of a person whose eyes have an accommodated power of 53.5 D?

Solution:

$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} \Rightarrow d_\text{o} = 28.6\,\text{cm}$

Exercise:

Problem:

(a) A laser reshaping the cornea of a myopic patient reduces the power of his eye by 9.00 D, with a $\pm 5.0\%$ uncertainty in the final correction. What is the range of diopters for eyeglass lenses that this person might need after this procedure? (b) Was the person nearsighted or farsighted before the procedure? How do you know?

Exercise:

Problem:

The power for normal close vision is 54.0 D. In a vision-correction procedure, the power of a patient's eye is increased by 3.00 D. Assuming that this produces normal close vision, what was the patient's near point before the procedure?

Solution:

Originally, the close vision was 51.0 D. Therefore, $P=\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}\Rightarrow d_{\text{o}}=1.00\,\text{m}$

Exercise:

Problem:

For normal distant vision, the eye has a power of 50.0 D. What was the previous far point of a patient who had laser vision correction that reduced the power of her eye by 7.00 D, producing normal distant vision?

Exercise:

Problem:

The power for normal distant vision is 50.0 D. A severely myopic patient has a far point of 5.00 cm. By how many diopters should the power of his eye be reduced in laser vision correction to obtain normal distant vision for him?

Solution:

originally, $P = 70.0\,\text{D}$; because the power for normal distant vision is 50.0 D, the power should be decreased by 20.0 D

Exercise:

Problem:

A student’s eyes, while reading the blackboard, have a power of 51.0 D. How far is the board from his eyes?

Exercise:

Problem:

The power of a physician’s eyes is 53.0 D while examining a patient. How far from her eyes is the object that is being examined?

Solution:

$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} \Rightarrow d_\text{o} = 0.333\,\text{m}$

Exercise:

Problem:

The normal power for distant vision is 50.0 D. A young woman with normal distant vision has a 10.0% ability to accommodate (that is, increase) the power of her eyes. What is the closest object she can see clearly?

Exercise:

Problem:

The far point of a myopic administrator is 50.0 cm. (a) What is the relaxed power of his eyes? (b) If he has the normal 8.00% ability to accommodate, what is the closest object he can see clearly?

Solution:

a. $P = 52.0\text{ D}$;
$P' = 56.16\text{ D}$

b. $\frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} = P \Rightarrow d_\text{o} = 16.2\text{ cm}$

Exercise:

Problem:

A very myopic man has a far point of 20.0 cm. What power contact lens (when on the eye) will correct his distant vision?

Exercise:

Problem:

Repeat the previous problem for eyeglasses held 1.50 cm from the eyes.

Solution:

We need $d_\text{i} = -18.5\text{ cm}$ when $d_\text{o} = \infty$, so
$P = -5.41\text{ D}$

Exercise:

Problem:

A myopic person sees that her contact lens prescription is –4.00 D. What is her far point?

Exercise:

Problem:

Repeat the previous problem for glasses that are 1.75 cm from the eyes.

Solution:

Let x = far point

$$\Rightarrow P = \frac{1}{-(x-0.0175\,\mathrm{m})} + \frac{1}{\infty} \Rightarrow -xP + (0.0175\,\mathrm{m})P = 1$$

$$\Rightarrow x = 26.8\,\mathrm{cm}$$

Exercise:

Problem:

The contact lens prescription for a mildly farsighted person is 0.750 D, and the person has a near point of 29.0 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Glossary

accommodation
use of the ciliary muscles to adjust the shape of the eye lens for focusing on near or far objects

far point
furthest point an eye can see in focus

farsightedness (or hyperopia)
visual defect in which near objects appear blurred because their images are focused behind the retina rather than on the retina; a farsighted person can see far objects clearly but near objects appear blurred

near point
closest point an eye can see in focus

nearsightedness (or myopia)
visual defect in which far objects appear blurred because their images are focused in front of the retina rather than on the retina; a nearsighted person can see near objects clearly but far objects appear blurred

optical power

(P) inverse of the focal length of a lens, with the focal length expressed in meters. The optical power P of a lens is expressed in units of diopters D; that is, $1\text{D} = 1/\text{m} = 1\,\text{m}^{-1}$

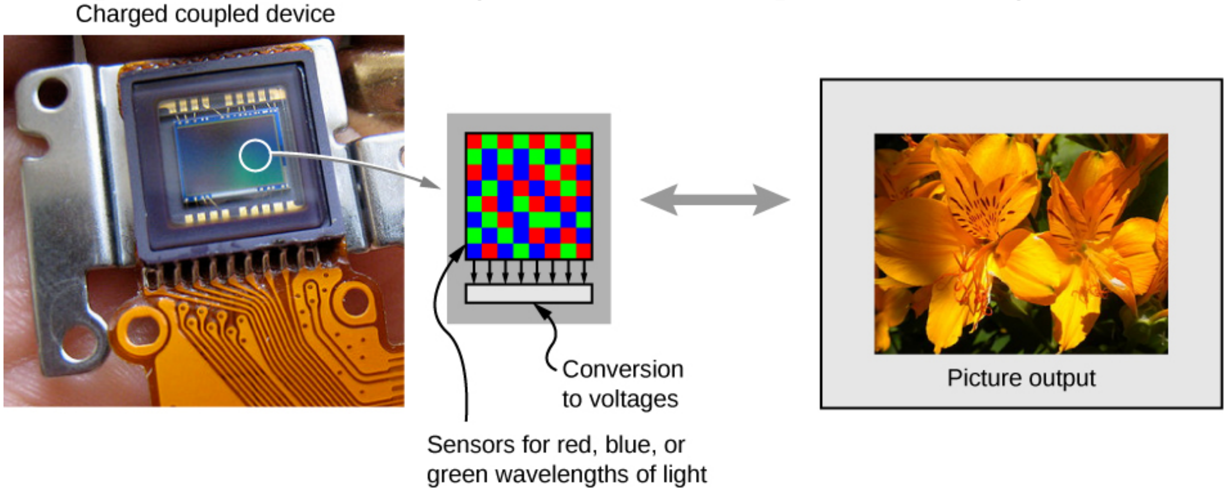
The Camera

By the end of this section, you will be able to:

- Describe the optics of a camera
- Characterize the image created by a camera

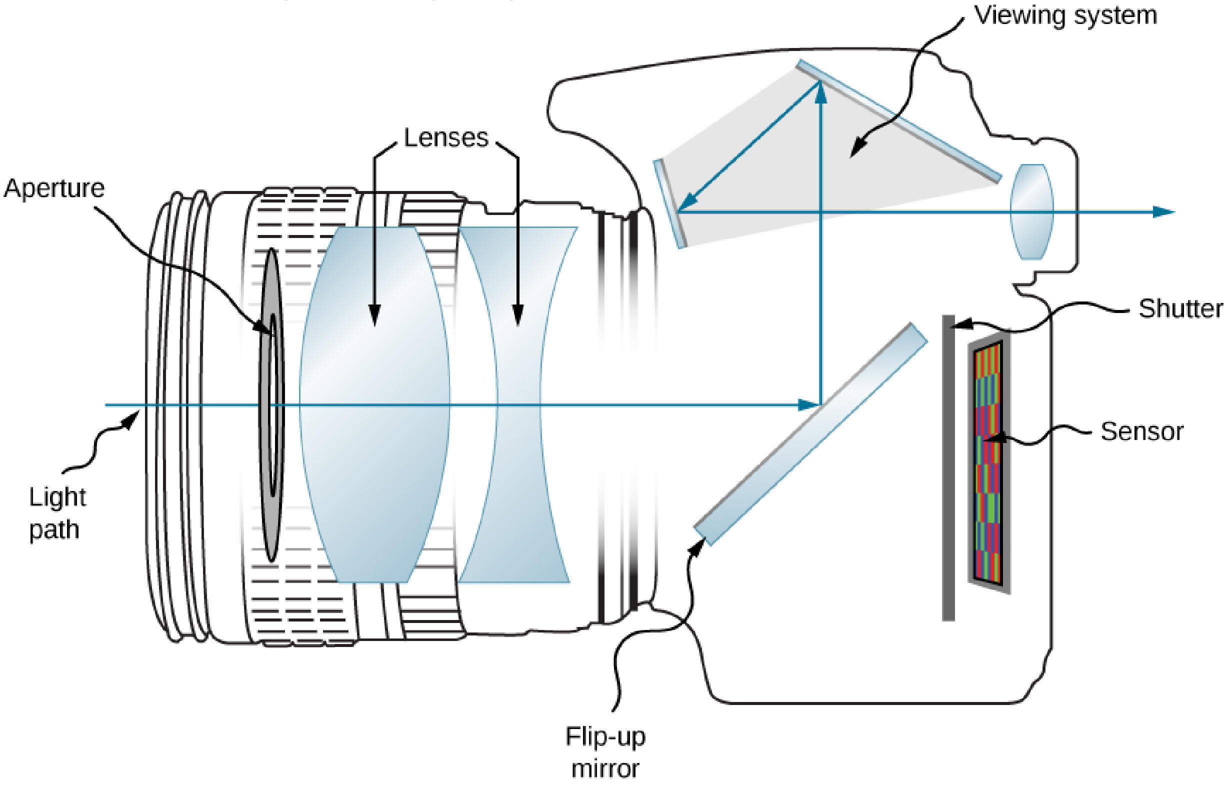
Cameras are very common in our everyday life. Between 1825 and 1827, French inventor Nicéphore Niépce successfully photographed an image created by a primitive camera. Since then, enormous progress has been achieved in the design of cameras and camera-based detectors.

Initially, photographs were recorded by using the light-sensitive reaction of silver-based compounds such as silver chloride or silver bromide. Silver-based photographic paper was in common use until the advent of digital photography in the 1980s, which is intimately connected to **charge-coupled device (CCD)** detectors. In a nutshell, a CCD is a semiconductor chip that records images as a matrix of tiny pixels, each pixel located in a “bin” in the surface. Each pixel is capable of detecting the intensity of light impinging on it. Color is brought into play by putting red-, blue-, and green-colored filters over the pixels, resulting in colored digital images ([link]). At its best resolution, one CCD pixel corresponds to one pixel of the image. To reduce the resolution and decrease the size of the file, we can “bin” several CCD pixels into one, resulting in a smaller but “pixelated” image.



A charge-coupled device (CCD) converts light signals into electronic signals, enabling electronic processing and storage of visual images. This is the basis for electronic imaging in all digital cameras, from cell phones to movie cameras. (credit left: modification of work by Bruce Turner)

Clearly, electronics is a big part of a digital camera; however, the underlying physics is basic optics. As a matter of fact, the optics of a camera are pretty much the same as those of a single lens with an object distance that is significantly larger than the lens's focal distance ([link]).



Modern digital cameras have several lenses to produce a clear image with minimal aberration and use red, blue, and green filters to produce a color image.

For instance, let us consider the camera in a smartphone. An average smartphone camera is equipped with a stationary wide-angle lens with a focal length of about 4–5 mm. (This focal length is about equal to the thickness of the phone.) The image created by the lens is focused on the

CCD detector mounted at the opposite side of the phone. In a cell phone, the lens and the CCD cannot move relative to each other. So how do we make sure that both the images of a distant and a close object are in focus?

Recall that a human eye can accommodate for distant and close images by changing its focal distance. A cell phone camera cannot do that because the distance from the lens to the detector is fixed. Here is where the small focal distance becomes important. Let us assume we have a camera with a 5-mm focal distance. What is the image distance for a selfie? The object distance for a selfie (the length of the hand holding the phone) is about 50 cm. Using the thin-lens equation, we can write

Equation:

$$\frac{1}{5\,\mathrm{mm}} = \frac{1}{500\,\mathrm{mm}} + \frac{1}{d_i}$$

We then obtain the image distance:

Equation:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{5\,\mathrm{mm}} - \frac{1}{500\,\mathrm{mm}}$$

Note that the object distance is 100 times larger than the focal distance. We can clearly see that the 1/(500 mm) term is significantly smaller than 1/(5 mm), which means that the image distance is pretty much equal to the lens's focal length. An actual calculation gives us the image distance $d_i = 5.05\,\mathrm{mm}$. This value is extremely close to the lens's focal distance.

Now let us consider the case of a distant object. Let us say that we would like to take a picture of a person standing about 5 m from us. Using the thin-lens equation again, we obtain the image distance of 5.005 mm. The farther the object is from the lens, the closer the image distance is to the focal distance. At the limiting case of an infinitely distant object, we obtain the image distance exactly equal to the focal distance of the lens.

As you can see, the difference between the image distance for a selfie and the image distance for a distant object is just about 0.05 mm or 50 microns. Even a short object distance such as the length of your hand is two orders of magnitude larger than the lens's focal length, resulting in minute variations of the image distance. (The 50-micron difference is smaller than the thickness of an average sheet of paper.) Such a small difference can be easily accommodated by the same detector, positioned at the focal distance of the lens. Image analysis software can help improve image quality.

Conventional point-and-shoot cameras often use a movable lens to change the lens-to-image distance. Complex lenses of the more expensive mirror reflex cameras allow for superb quality photographic images. The optics of these camera lenses is beyond the scope of this textbook.

Summary

- Cameras use combinations of lenses to create an image for recording.
- Digital photography is based on charge-coupled devices (CCDs) that break an image into tiny "pixels" that can be converted into electronic signals.

Glossary

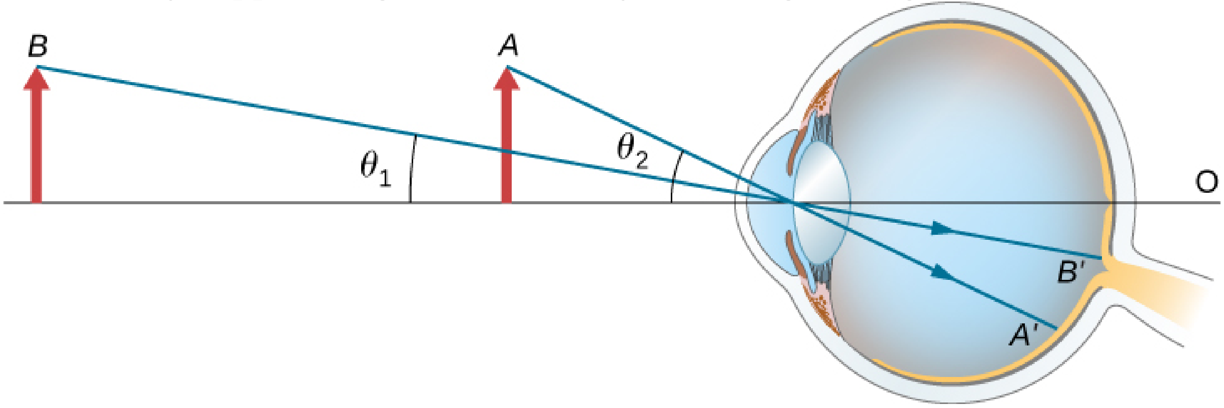
charge-coupled device (CCD)
semiconductor chip that converts a light image into tiny pixels that can be converted into electronic signals of color and intensity

The Simple Magnifier

By the end of this section, you will be able to:

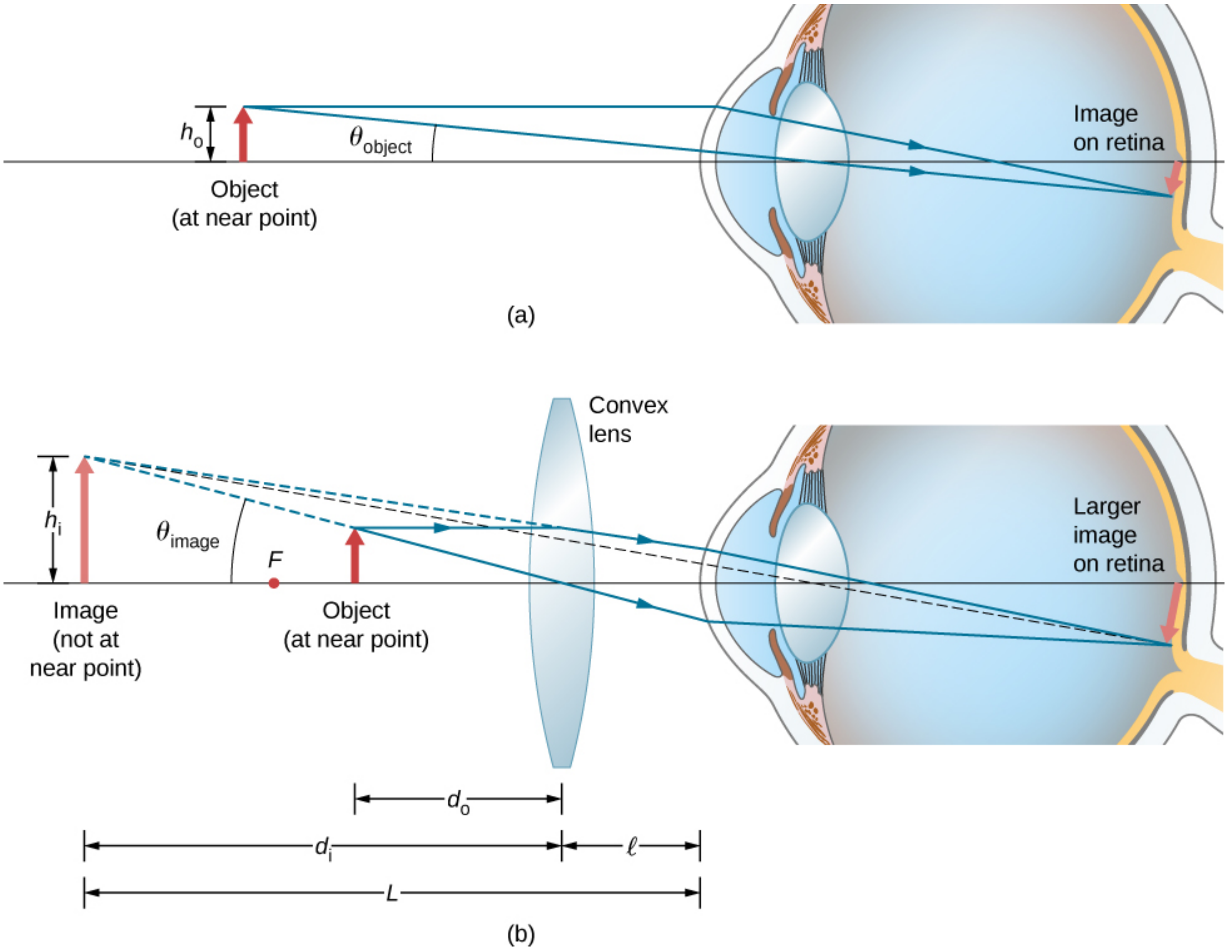
- Understand the optics of a simple magnifier
- Characterize the image created by a simple magnifier

The apparent size of an object perceived by the eye depends on the angle the object subtends from the eye. As shown in [link], the object at A subtends a larger angle from the eye than when it is position at point B. Thus, the object at A forms a larger image on the retina (see OA') than when it is positioned at B (see OB'). Thus, objects that subtend large angles from the eye appear larger because they form larger images on the retina.



Size perceived by an eye is determined by the angle subtended by the object. An image formed on the retina by an object at A is larger than an image formed on the retina by the same object positioned at B (compared image heights OA' to OB').

We have seen that, when an object is placed within a focal length of a convex lens, its image is virtual, upright, and larger than the object (see part (b) of [link]). Thus, when such an image produced by a convex lens serves as the object for the eye, as shown in [link], the image on the retina is enlarged, because the image produced by the lens subtends a larger angle in the eye than does the object. A convex lens used for this purpose is called a **magnifying glass** or a **simple magnifier**.


h_o
θ_{object}
Object
(at near point)
Image
on retina
(a)
Convex
lens
h_i
θ_{image}
F
Image
(not at
near point)
Object
(at near point)
Larger
image
on retina
d_o
d_i
ℓ
L
(b)

The simple magnifier is a convex lens used to produce an enlarged image of an object on the retina. (a) With no convex lens, the object subtends an angle θ_{object} from the eye. (b) With the convex lens in place, the image produced by the convex lens subtends an angle θ_{image} from the eye, with $\theta_{\text{image}} > \theta_{\text{object}}$. Thus, the image on the retina is larger with the convex lens in place.

To account for the magnification of a magnifying lens, we compare the angle subtended by the image (created by the lens) with the angle subtended by the object (viewed with no lens), as shown in [link]. We assume that the object is situated at the near point of the eye, because this is the object distance at which the unaided eye can form the largest image on the retina. We will compare the magnified images created by a lens with this maximum image size for the unaided eye. The magnification of an image when observed by the eye is the **angular magnification** M, which is defined by the ratio of the angle θ_{image} subtended by the image to the angle θ_{object} subtended by the object:

Note:
Equation:

$$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}}.$$

Consider the situation shown in [link]. The magnifying lens is held a distance ℓ from the eye, and the image produced by the magnifier forms a distance L from the eye. We want to calculate the angular magnification for any arbitrary L and ℓ. In the small-angle approximation, the angular size θ_{image} of the image is h_{i}/L. The angular size θ_{object} of the object at the near point is $\theta_{\text{object}} = h_{\text{o}}/25\text{ cm}$. The angular magnification is then

Equation:

$$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}} = \frac{h_{\text{i}}(25\text{ cm})}{Lh_{\text{o}}}.$$

Using [link] for linear magnification

Equation:

$$m = -\frac{d_{\text{i}}}{d_{\text{o}}} = \frac{h_{\text{i}}}{h_{\text{o}}}$$

and the thin-lens equation

Equation:

$$\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}=\frac{1}{f}$$

in [link], we arrive at the following expression for the angular magnification of a magnifying lens:

Equation:

$$\begin{aligned} M &= \left(-\tfrac{d_\text{i}}{d_\text{o}}\right)\left(\tfrac{25\ \text{cm}}{L}\right) \\ &= -d_\text{i}\left(\tfrac{1}{f}-\tfrac{1}{d_\text{i}}\right)\left(\tfrac{25\ \text{cm}}{L}\right) \\ &= \left(1-\tfrac{d_\text{i}}{f}\right)\left(\tfrac{25\ \text{cm}}{L}\right) \end{aligned}$$

From part (b) of the figure, we see that the absolute value of the image distance is $|d_\text{i}|=L-\ell$. Note that $d_\text{i}<0$ because the image is virtual, so we can dispense with the absolute value by explicitly inserting the minus sign: $-d_\text{i}=L-\ell$. Inserting this into [link] gives us the final equation for the angular magnification of a magnifying lens:

Note:
Equation:

$$M=\left(\frac{25\ \text{cm}}{L}\right)\left(1+\frac{L-\ell}{f}\right).$$

Note that all the quantities in this equation have to be expressed in centimeters. Often, we want the image to be at the near-point distance ($L=25\ \text{cm}$) to get maximum magnification, and we hold the magnifying lens close to the eye ($\ell=0$). In this case, [link] gives

Equation:

$$M = 1 + \frac{25\ \text{cm}}{f}$$

which shows that the greatest magnification occurs for the lens with the shortest focal length. In addition, when the image is at the near-point distance and the lens is held close to the eye $(\ell = 0)$, then $L = d_\text{i} = 25\ \text{cm}$ and [link] becomes

Equation:

$$M = \frac{h_\text{i}}{h_\text{o}} = m$$

where *m* is the linear magnification ([link]) derived for spherical mirrors and thin lenses. Another useful situation is when the image is at infinity $(L = \infty)$. [link] then takes the form

Equation:

$$M(L = \infty) = \frac{25\ \text{cm}}{f}.$$

The resulting magnification is simply the ratio of the near-point distance to the focal length of the magnifying lens, so a lens with a shorter focal length gives a stronger magnification. Although this magnification is smaller by 1 than the magnification obtained with the image at the near point, it provides for the most comfortable viewing conditions, because the eye is relaxed when viewing a distant object.

By comparing [link] with [link], we see that the range of angular magnification of a given converging lens is

Note:

Equation:

$$\frac{25\text{ cm}}{f} \leq M \leq 1 + \frac{25\text{ cm}}{f}.$$

Example:

Magnifying a Diamond

A jeweler wishes to inspect a 3.0-mm-diameter diamond with a magnifier. The diamond is held at the jeweler’s near point (25 cm), and the jeweler holds the magnifying lens close to his eye.

(a) What should the focal length of the magnifying lens be to see a 15-mm-diameter image of the diamond?

(b) What should the focal length of the magnifying lens be to obtain 10 × magnification?

Strategy

We need to determine the requisite magnification of the magnifier. Because the jeweler holds the magnifying lens close to his eye, we can use [link] to find the focal length of the magnifying lens.

Solution

a. The required linear magnification is the ratio of the desired image diameter to the diamond’s actual diameter ([link]). Because the jeweler holds the magnifying lens close to his eye and the image forms at his near point, the linear magnification is the same as the angular magnification, so

   Equation:

$$M = m = \frac{h_\text{i}}{h_\text{o}} = \frac{15\text{ mm}}{3.0\text{ mm}} = 5.0.$$

   The focal length *f* of the magnifying lens may be calculated by solving [link] for *f*, which gives

   Equation:

$$\begin{aligned} M &= 1 + \frac{25\text{ cm}}{f} \\ f &= \frac{25\text{ cm}}{M-1} = \frac{25\text{ cm}}{5.0-1} = 6.3\text{ cm} \end{aligned}$$

b. To get an image magnified by a factor of ten, we again solve [link] for *f*, but this time we use $M = 10$. The result is

Equation:

$$f = \frac{25\text{ cm}}{M-1} = \frac{25\text{ cm}}{10-1} = 2.8\text{ cm}.$$

Significance

Note that a greater magnification is achieved by using a lens with a smaller focal length. We thus need to use a lens with radii of curvature that are less than a few centimeters and hold it very close to our eye. This is not very convenient. A compound microscope, explored in the following section, can overcome this drawback.

Summary

- A simple magnifier is a converging lens and produces a magnified virtual image of an object located within the focal length of the lens.
- Angular magnification accounts for magnification of an image created by a magnifier. It is equal to the ratio of the angle subtended by the image to that subtended by the object when the object is observed by the unaided eye.
- Angular magnification is greater for magnifying lenses with smaller focal lengths.
- Simple magnifiers can produce as great as tenfold ($10\,\times$) magnification.

Problems

Exercise:

Problem:

If the image formed on the retina subtends an angle of 30° and the object subtends an angle of 5°, what is the magnification of the image?

Solution:

$M = 6\ \times$

Exercise:

Problem:

What is the magnification of a magnifying lens with a focal length of 10 cm if it is held 3.0 cm from the eye and the object is 12 cm from the eye?

Exercise:

Problem:

How far should you hold a 2.1 cm-focal length magnifying glass from an object to obtain a magnification of $10\ \times$? Assume you place your eye 5.0 cm from the magnifying glass.

Solution:

$$\begin{aligned} M &= \left(\frac{25\text{ cm}}{L}\right)\left(1+\frac{L-\ell}{f}\right) \\ L-\ell &= d_{\text{o}} \\ d_{\text{o}} &= 13\text{ cm} \end{aligned}$$

Exercise:

Problem:

You hold a 5.0 cm-focal length magnifying glass as close as possible to your eye. If you have a normal near point, what is the magnification?

Exercise:

Problem:

You view a mountain with a magnifying glass of focal length $f = 10\text{ cm}$. What is the magnification?

Solution:

$M = 2.5\ \times$

Exercise:

Problem:

You view an object by holding a 2.5 cm-focal length magnifying glass 10 cm away from it. How far from your eye should you hold the magnifying glass to obtain a magnification of $10 \times$?

Exercise:

Problem:

A magnifying glass forms an image 10 cm on the opposite side of the lens from the object, which is 10 cm away. What is the magnification of this lens for a person with a normal near point if their eye 12 cm from the object?

Solution:

$M \quad = -2.1 \times$

Exercise:

Problem:

An object viewed with the naked eye subtends a 2° angle. If you view the object through a $10\ \times$ magnifying glass, what angle is subtended by the image formed on your retina?

Exercise:

Problem:

For a normal, relaxed eye, a magnifying glass produces an angular magnification of 4.0. What is the largest magnification possible with this magnifying glass?

Solution:

$$M = \frac{25\text{ cm}}{f}$$

$$M_{\text{max}} = 5$$

Exercise:

Problem:

What range of magnification is possible with a 7.0 cm-focal length converging lens?

Exercise:

Problem:

A magnifying glass produces an angular magnification of 4.5 when used by a young person with a near point of 18 cm. What is the maximum angular magnification obtained by an older person with a near point of 45 cm?

Solution:

$$\begin{aligned} M_{\text{max}}^{\text{young}} &= 1 + \frac{18\text{ cm}}{f} \Rightarrow f = \frac{18\text{ cm}}{M_{\text{max}}^{\text{young}} - 1} \\ M_{\text{max}}^{\text{old}} &= 9.8\times \end{aligned}$$

Glossary

angular magnification
: ratio of the angle subtended by an object observed with a magnifier to that observed by the naked eye

simple magnifier (or magnifying glass)

converging lens that produces a virtual image of an object that is within the focal length of the lens

Microscopes and Telescopes

By the end of this section, you will be able to:

- Explain the physics behind the operation of microscopes and telescopes
- Describe the image created by these instruments and calculate their magnifications

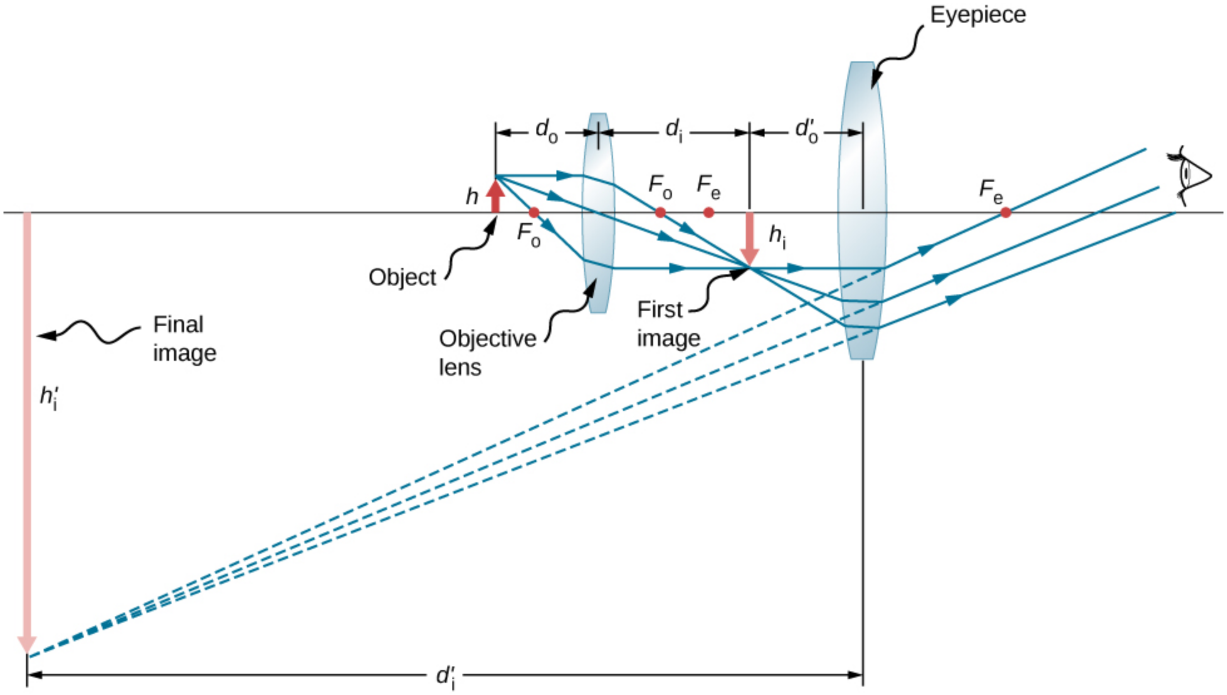
Microscopes and telescopes are major instruments that have contributed hugely to our current understanding of the micro- and macroscopic worlds. The invention of these devices led to numerous discoveries in disciplines such as physics, astronomy, and biology, to name a few. In this section, we explain the basic physics that make these instruments work.

Microscopes

Although the eye is marvelous in its ability to see objects large and small, it obviously is limited in the smallest details it can detect. The desire to see beyond what is possible with the naked eye led to the use of optical instruments. We have seen that a simple convex lens can create a magnified image, but it is hard to get large magnification with such a lens. A magnification greater than $5\times$ is difficult without distorting the image. To get higher magnification, we can combine the simple magnifying glass with one or more additional lenses. In this section, we examine microscopes that enlarge the details that we cannot see with the naked eye.

Microscopes were first developed in the early 1600s by eyeglass makers in The Netherlands and Denmark. The simplest **compound microscope** is constructed from two convex lenses ([link]). The **objective** lens is a convex lens of short focal length (i.e., high power) with typical magnification from $5\times$ to $100\times$. The **eyepiece**, also referred to as the ocular, is a convex lens of longer focal length.

The purpose of a microscope is to create magnified images of small objects, and both lenses contribute to the final magnification. Also, the final enlarged image is produced sufficiently far from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close (i.e., closer than the near point of the eye).



A compound microscope is composed of two lenses: an objective and an eyepiece. The objective forms the first image, which is larger than the object. This first image is inside the focal length of the eyepiece and serves as the object for the eyepiece. The eyepiece forms final image that is further magnified.

To see how the microscope in [link] forms an image, consider its two lenses in succession. The object is just beyond the focal length f^{obj} of the objective lens, producing a real, inverted image that is larger than the object. This first image serves as the object for the second lens, or eyepiece. The eyepiece is positioned so that the first image is within its focal length f^{eye}, so that it can further magnify the image. In a sense, it acts as a magnifying glass that magnifies the intermediate image produced by the objective. The image produced by the eyepiece is a magnified virtual image. The final image remains inverted but is farther from the observer than the object, making it easy to view.

The eye views the virtual image created by the eyepiece, which serves as the object for the lens in the eye. The virtual image formed by the eyepiece is well outside the focal length of the eye, so the eye forms a real image on the retina.

The magnification of the microscope is the product of the linear magnification m^{obj} by the objective and the angular magnification M^{eye} by the eyepiece. These are given by

Equation:

$$\begin{aligned} m^{\text{obj}} &= -\frac{d_{\text{i}}^{\text{obj}}}{d_{\text{o}}^{\text{obj}}} \approx -\frac{d_{\text{i}}^{\text{obj}}}{f^{\text{obj}}} \text{ (linear magnification by objective)} \\ M^{\text{eye}} &= 1 + \frac{25\text{ cm}}{f^{\text{eye}}} \text{ (angular magnification by eyepiece)} \end{aligned}$$

Here, f^{obj} and f^{eye} are the focal lengths of the objective and the eyepiece, respectively. We assume that the final image is formed at the near point of the eye, providing the largest magnification. Note that the angular magnification of the eyepiece is the same as obtained earlier for the simple magnifying glass. This should not be surprising, because the eyepiece is essentially a magnifying glass, and the same physics applies here. The **net magnification** M_{net} of the compound microscope is the product of the linear magnification of the objective and the angular magnification of the eyepiece:

Note:

Equation:

$$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{d_{\text{i}}^{\text{obj}} \left(f^{\text{eye}} + 25\text{ cm}\right)}{f^{\text{obj}} f^{\text{eye}}}.$$

Example:

Microscope Magnification

Calculate the magnification of an object placed 6.20 mm from a compound microscope that has a 6.00 mm-focal length objective and a 50.0 mm-focal length eyepiece. The objective and eyepiece are separated by 23.0 cm.

Strategy

This situation is similar to that shown in [link]. To find the overall magnification, we must know the linear magnification of the objective and the angular magnification of the eyepiece. We can use [link], but we need to use the thin-lens equation to find the image distance $d_\mathrm{i}^\mathrm{obj}$ of the objective.

Solution

Solving the thin-lens equation for $d_\mathrm{i}^\mathrm{obj}$ gives

Equation:

$$\begin{aligned} d_\mathrm{i}^\mathrm{obj} &= \left(\frac{1}{f^\mathrm{obj}} - \frac{1}{d_\mathrm{o}^\mathrm{obj}}\right)^{-1} \\ &= \left(\frac{1}{6.00\text{ mm}} - \frac{1}{6.20\text{ mm}}\right)^{-1} = 186\text{ mm} = 18.6\text{ cm} \end{aligned}$$

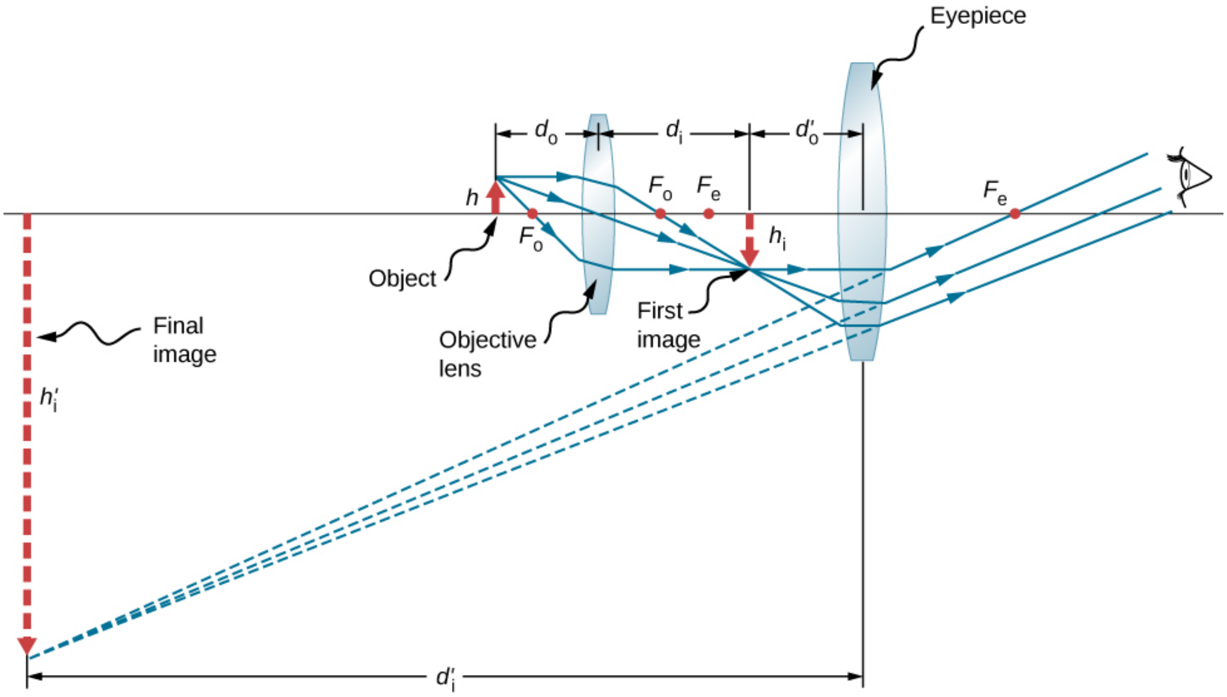
Inserting this result into [link] along with the known values $f^\mathrm{obj} = 6.00\text{ mm} = 0.600\text{ cm}$ and $f^\mathrm{eye} = 50.0\text{ mm} = 5.00\text{ cm}$ gives

Equation:

$$\begin{aligned} M_\mathrm{net} &= -\frac{d_\mathrm{i}^\mathrm{obj}(f^\mathrm{eye} + 25\text{ cm})}{f^\mathrm{obj} f^\mathrm{eye}} \\ &= -\frac{(18.6\text{ cm})(5.00\text{ cm} + 25\text{ cm})}{(0.600\text{ cm})(5.00\text{ cm})} \\ &= -186 \end{aligned}$$

Significance

Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with [link], where the image is seen to be large and inverted. In this case, the image is virtual and inverted, which cannot happen for a single element (see [link]).



A compound microscope with the image created at infinity.

We now calculate the magnifying power of a microscope when the image is at infinity, as shown in [link], because this makes for the most relaxed viewing. The magnifying power of the microscope is the product of linear magnification m^{obj} of the objective and the angular magnification M^{eye} of the eyepiece. We know that $m^{\text{obj}} = -d_{\text{i}}^{\text{obj}}/d_{\text{o}}^{\text{obj}}$ and from the thin-lens equation we obtain

Equation:

$$m^{\text{obj}} = -\frac{d_{\text{i}}^{\text{obj}}}{d_{\text{o}}^{\text{obj}}} = 1 - \frac{d_{\text{i}}^{\text{obj}}}{f^{\text{obj}}} = \frac{f^{\text{obj}} - d_{\text{i}}^{\text{obj}}}{f^{\text{obj}}}.$$

If the final image is at infinity, then the image created by the objective must be located at the focal point of the eyepiece. This may be seen by considering the thin-lens equation with $d_{\text{i}} = \infty$ or by recalling that rays that pass through the focal point exit the lens parallel to each other, which is equivalent to focusing at infinity. For many microscopes, the distance between the image-side focal point of the objective and the object-side focal point of the eyepiece is standardized at $L = 16\text{ cm}$. This distance is called the tube length of the microscope. From [link], we see that $L = f^{\text{obj}} - d_{\text{i}}^{\text{obj}}$. Inserting this into [link] gives

Equation:

$$m^{\text{obj}} = \frac{L}{f^{\text{obj}}} = \frac{16\text{ cm}}{f^{\text{obj}}}.$$

We now need to calculate the angular magnification of the eyepiece with the image at infinity. To do so, we take the ratio of the angle θ_{image} subtended by the image to the angle θ_{object} subtended by the object at the near point of the eye (this is the closest that the unaided eye can view the object, and thus this is the position where the object will form the largest image on the retina of the unaided eye). Using [link] and working in the small-angle approximation, we have $\theta_{\text{image}} \approx h_{\text{i}}^{\text{obj}}/f^{\text{eye}}$ and $\theta_{\text{object}} \approx h_{\text{i}}^{\text{obj}}/25\text{ cm}$, where $h_{\text{i}}^{\text{obj}}$ is the height of the image formed by the objective, which is the object of the eyepiece. Thus, the angular magnification of the eyepiece is

Equation:

$$M^{\text{eye}} = \frac{\theta_{\text{image}}}{\theta_{\text{object}}} = \frac{h_{\text{i}}^{\text{obj}}}{f^{\text{eye}}} \frac{25\text{ cm}}{h_{\text{i}}^{\text{obj}}} = \frac{25\text{ cm}}{f^{\text{eye}}}.$$

The net magnifying power of the compound microscope with the image at infinity is therefore

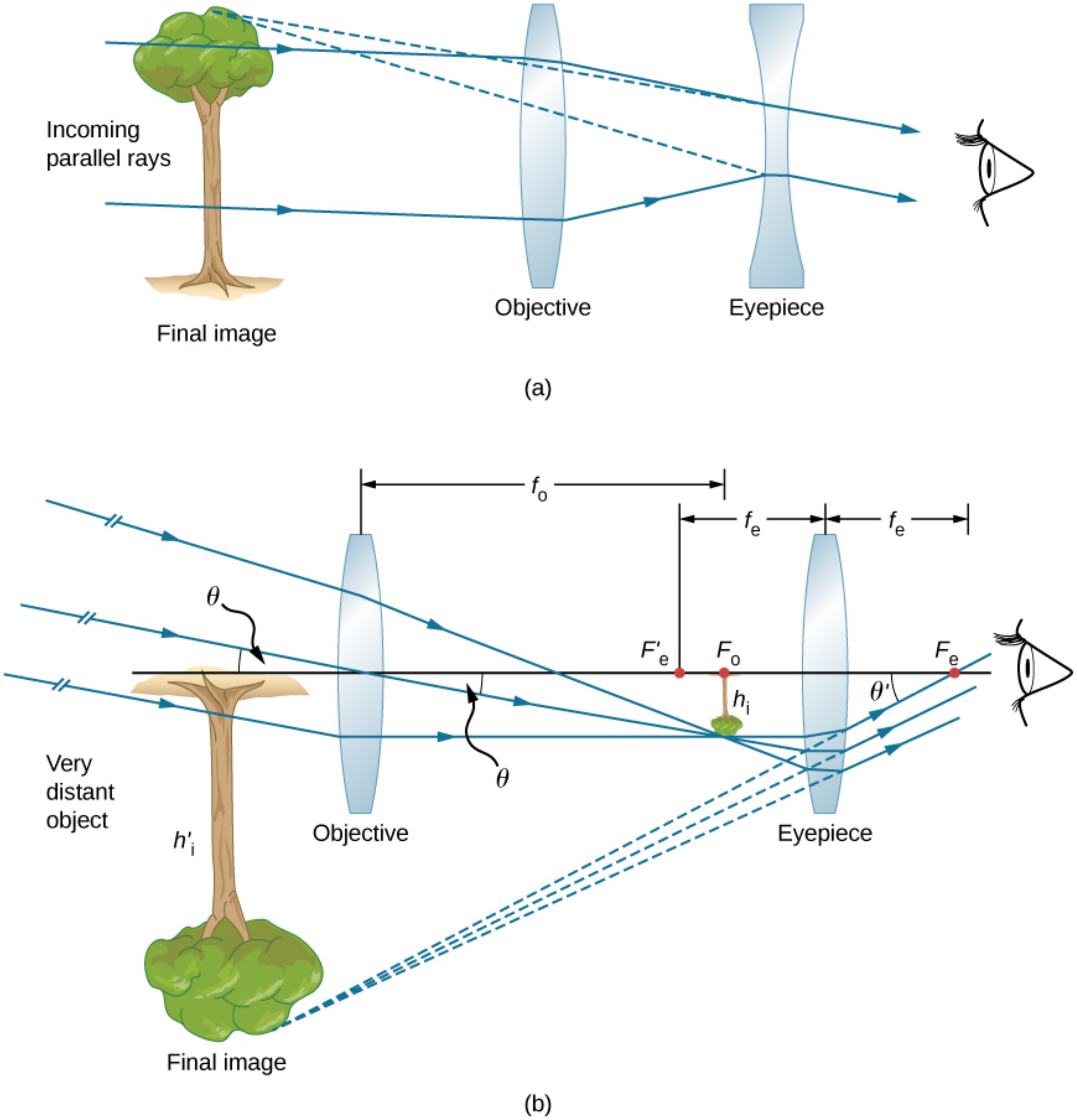
Equation:

$$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{(16\text{ cm})(25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}}.$$

The focal distances must be in centimeters. The minus sign indicates that the final image is inverted. Note that the only variables in the equation are the focal distances of the eyepiece and the objective, which makes this equation particularly useful.

Telescopes

Telescopes are meant for viewing distant objects and produce an image that is larger than the image produced in the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Telescopes were invented around 1600, and Galileo was the first to use them to study the heavens, with monumental consequences. He observed the moons of Jupiter, the craters and mountains on the moon, the details of sunspots, and the fact that the Milky Way is composed of a vast number of individual stars.


Incoming parallel rays
Final image
Objective
Eyepiece
(a)
f_o
f_e
f_e
θ
F'_e
F_o
F_e
h_i
θ'
θ
Very distant object
h'_i
Objective
Eyepiece
Final image
(b)

(a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple refracting telescopes have two convex lenses. The objective forms a real, inverted image at (or just within) the focal plane of the eyepiece. This image serves as the object for the eyepiece. The eyepiece forms a virtual, inverted image that is magnified.

Part (a) of [link] shows a refracting telescope made of two lenses. The first lens, called the objective, forms a real image within the focal length of the second lens, which is called the eyepiece. The image of the objective lens serves as the object for the eyepiece, which forms a magnified virtual image that is observed by the eye. This design is what Galileo used to observe the heavens.

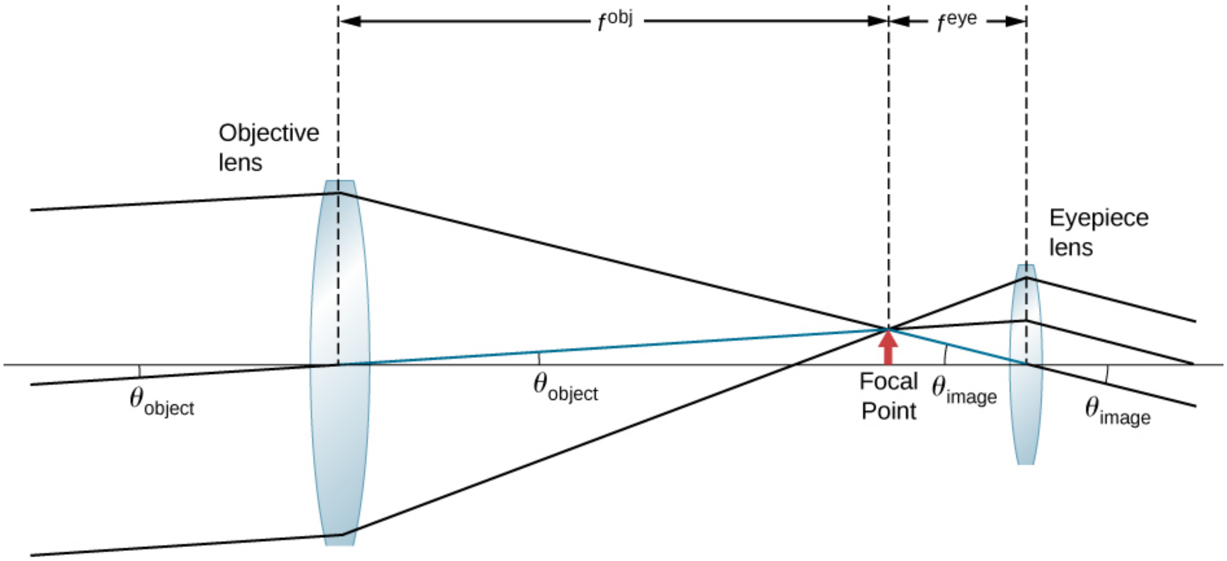
Although the arrangement of the lenses in a refracting telescope looks similar to that in a microscope, there are important differences. In a telescope, the real object is far away and the intermediate image is smaller than the object. In a microscope, the real object is very close and the intermediate image is larger than the object. In both the telescope and the microscope, the eyepiece magnifies the intermediate image; in the telescope, however, this is the only magnification.

The most common two-lens telescope is shown in part (b) of the figure. The object is so far from the telescope that it is essentially at infinity compared with the focal lengths of the lenses $\left(d_{\mathrm{o}}^{\mathrm{obj}} \approx \infty\right)$, so the incoming rays are essentially parallel and focus on the focal plane. Thus, the first image is produced at $d_{\mathrm{i}}^{\mathrm{obj}} = f^{\mathrm{obj}}$, as shown in the figure, and is not large compared with what you might see by looking directly at the object. However, the eyepiece of the telescope eyepiece (like the microscope eyepiece) allows you to get nearer than your near point to this first image and so magnifies it (because you are near to it, it subtends a larger angle from your eye and so forms a larger image on your retina). As for a simple magnifier, the angular magnification of a telescope is the ratio of the angle subtended by the image [θ_{image} in part (b)] to the angle subtended by the real object [θ_{object} in part (b)]:

Equation:

$$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}}.$$

To obtain an expression for the magnification that involves only the lens parameters, note that the focal plane of the objective lens lies very close to the focal plan of the eyepiece. If we assume that these planes are superposed, we have the situation shown in [link].



f^{obj}
f^{eye}
Objective lens
Eyepiece lens
θ_{object}
θ_{object}
Focal Point
θ_{image}
θ_{image}

The focal plane of the objective lens of a telescope is very near to the focal plane of the eyepiece. The angle θ_{image} subtended by the image viewed through the eyepiece is larger than the angle θ_{object} subtended by the object when viewed with the unaided eye.

We further assume that the angles θ_{object} and θ_{image} are small, so that the small-angle approximation holds ($\tan\theta \approx \theta$). If the image formed at the focal plane has height *h*, then

Equation:

$$\theta_{\text{object}} \approx \tan\theta_{\text{object}} = \frac{h}{f^{\text{obj}}}$$

$$\theta_{\text{image}} \approx \tan\theta_{\text{image}} = \frac{-h}{f^{\text{eye}}}$$

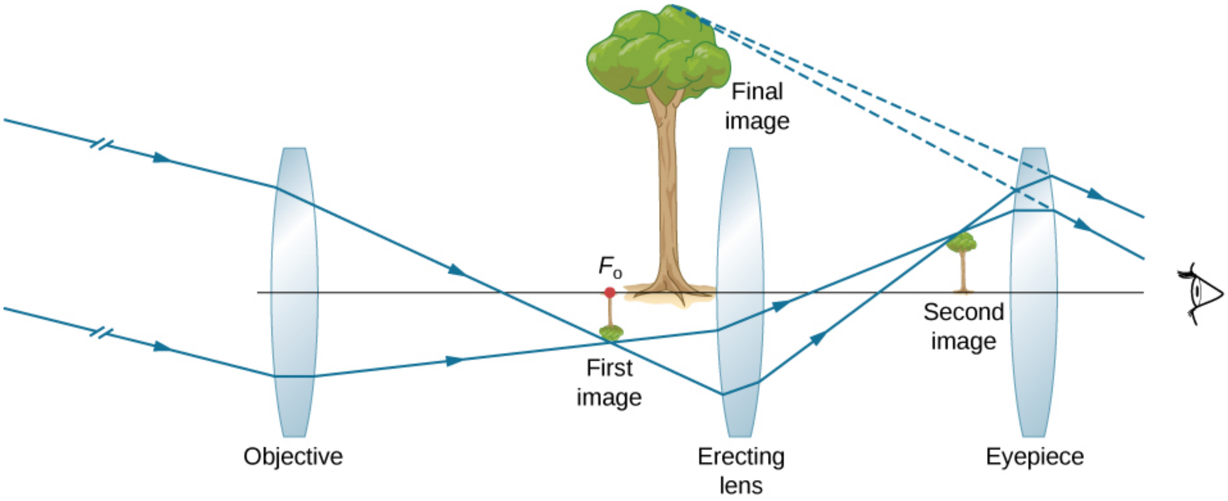
where the minus sign is introduced because the height is negative if we measure both angles in the counterclockwise direction. Inserting these expressions into [link] gives

Equation:

$$M = \frac{-h_{\text{i}}}{f^{\text{eye}}} \frac{f^{\text{obj}}}{h_{\text{i}}} = -\frac{f^{\text{obj}}}{f^{\text{eye}}}.$$

Thus, to obtain the greatest angular magnification, it is best to have an objective with a long focal length and an eyepiece with a short focal length. The greater the angular magnification *M*, the larger an object will appear when viewed through a telescope, making more details visible. Limits to observable details are imposed by many factors, including lens quality and atmospheric disturbance. Typical eyepieces have focal lengths of 2.5 cm or 1.25 cm. If the objective of the telescope has a focal length of 1 meter, then these eyepieces result in magnifications of 40 × and 80 × , respectively. Thus, the angular magnifications make the image appear 40 times or 80 times closer than the real object.

The minus sign in the magnification indicates the image is inverted, which is unimportant for observing the stars but is a real problem for other applications, such as telescopes on ships or telescopic gun sights. If an upright image is needed, Galileo’s arrangement in part (a) of [link] can be used. But a more common arrangement is to use a third convex lens as an eyepiece, increasing the distance between the first two and inverting the image once again, as seen in [link].



This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

The largest refracting telescope in the world is the 40-inch diameter Yerkes telescope located at Lake Geneva, Wisconsin ([link]), and operated by the University of Chicago.

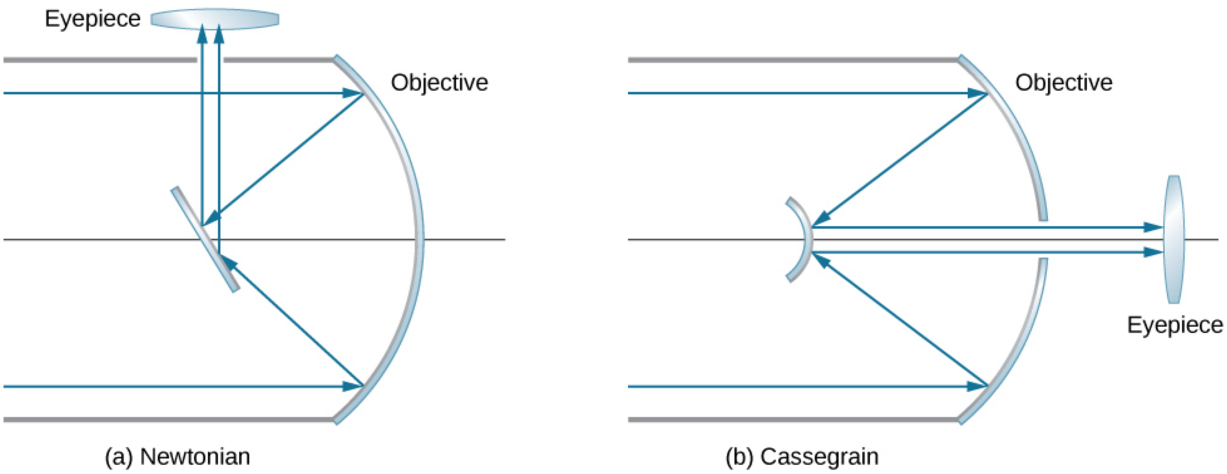
It is very difficult and expensive to build large refracting telescopes. You need large defect-free lenses, which in itself is a technically demanding task. A refracting telescope basically looks like a tube with a support structure to rotate it in different directions. A refracting telescope suffers from several problems. The aberration of lenses causes the image to be blurred. Also, as the lenses become thicker for larger lenses, more light is absorbed, making faint stars more difficult to observe. Large lenses are also very heavy and deform under their own weight. Some of these problems with refracting telescopes are addressed by avoiding refraction for collecting light and instead using a curved mirror in its place, as devised by Isaac Newton. These telescopes are called reflecting telescopes.

In 1897, the Yerkes Observatory in Wisconsin (USA) built a large refracting telescope with an objective lens that is 40 inches in diameter and has a tube length of 62 feet. (credit: Yerkes Observatory, University of Chicago)

Reflecting Telescopes

Isaac Newton designed the first reflecting telescope around 1670 to solve the problem of chromatic aberration that happens in all refracting telescopes. In chromatic aberration, light of different colors refracts by slightly different amounts in the lens. As a result, a rainbow appears around the image and the image appears blurred. In the reflecting telescope, light rays from a distant source fall upon the surface of a concave mirror fixed at the bottom end of the tube. The use of a mirror instead of a lens eliminates chromatic aberration. The concave mirror focuses the rays on its focal plane. The design problem is how to observe the focused image. Newton used a design in which the focused light from the concave mirror was reflected to one side of the tube into an eyepiece [part (a) of [link]]. This arrangement is common in many amateur telescopes and is called the **Newtonian design**.

Some telescopes reflect the light back toward the middle of the concave mirror using a convex mirror. In this arrangement, the light-gathering concave mirror has a hole in the middle [part (b) of the figure]. The light then is incident on an eyepiece lens. This arrangement of the objective and eyepiece is called the **Cassegrain design**. Most big telescopes, including the Hubble space telescope, are of this design. Other arrangements are also possible. In some telescopes, a light detector is placed right at the spot where light is focused by the curved mirror.



Reflecting telescopes: (a) In the Newtonian design, the eyepiece is located at the side of the telescope; (b) in the Cassegrain design, the eyepiece is located past a hole in the primary mirror.

Most astronomical research telescopes are now of the reflecting type. One of the earliest large telescopes of this kind is the Hale 200-inch (or 5-meter) telescope built on Mount Palomar in southern California, which has a 200 inch-diameter mirror. One of the largest telescopes in the world is the 10-meter Keck telescope at the Keck Observatory on the summit of the dormant Mauna Kea volcano in Hawaii. The Keck Observatory operates two 10-meter telescopes. Each is not a single mirror, but is instead made up of 36 hexagonal mirrors. Furthermore, the two telescopes on the Keck can work together, which increases their power to an effective 85-meter mirror. The Hubble telescope ([link]) is another large reflecting telescope with a 2.4 meter-diameter primary mirror. The Hubble was put into orbit around Earth in 1990.

The Hubble space telescope as seen from the Space Shuttle Discovery. (credit: modification of work by NASA)

The angular magnification *M* of a reflecting telescope is also given by [link]. For a spherical mirror, the focal length is half the radius of curvature, so making a large objective mirror not only helps the telescope collect more light but also increases the magnification of the image.

Summary

- Many optical devices contain more than a single lens or mirror. These are analyzed by considering each element sequentially. The image formed by the first is the object for the second, and so on. The same ray-tracing and thin-lens techniques developed in the previous sections apply to each lens element.
- The overall magnification of a multiple-element system is the product of the linear magnifications of its individual elements times the angular magnification of the eyepiece. For a two-element system with an objective and an eyepiece, this is
  Equation:

  $$M = m^{\text{obj}} M^{\text{eye}}.$$

  where m^{obj} is the linear magnification of the objective and M^{eye} is the angular magnification of the eyepiece.
- The microscope is a multiple-element system that contains more than a single lens or mirror. It allows us to see detail that we could not to see with the unaided eye. Both the eyepiece and objective contribute to the magnification. The magnification of a compound microscope with the image at infinity is
  Equation:

  $$M_{\text{net}} = -\frac{(16\text{ cm})(25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}}.$$

  In this equation, 16 cm is the standardized distance between the image-side focal point of the objective lens and the object-side focal point of the eyepiece, 25 cm is the normal near point distance, f^{obj} and f^{eye} are the focal distances for the objective lens and the eyepiece, respectively.

- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances.
- The angular magnification M for a telescope is given by
  Equation:

  $$M=-\frac{f^{\text{obj}}}{f^{\text{eye}}},$$

  where f^{obj} and f^{eye} are the focal lengths of the objective lens and the eyepiece, respectively.

Key Equations

Image distance in a plane mirror	$d_{\text{o}}=-d_{\text{i}}$
Focal length for a spherical mirror	$f=\frac{R}{2}$
Mirror equation	$\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}=\frac{1}{f}$
Magnification of a spherical mirror	$m=\frac{h_{\text{i}}}{h_{\text{o}}}=-\frac{d_{\text{i}}}{d_{\text{o}}}$
Sign convention for mirrors	
Focal length f	$+$ for concave mirror $-$ for convex mirror

Object distance d_o	$+$ for real object $-$ for virtual object
Image distance d_i	$+$ for real image $-$ for virtual image
Magnification m	$+$ for upright image $-$ for inverted image
Apparent depth equation	$h_\text{i} = \left(\frac{n_2}{n_1}\right) h_\text{o}$
Spherical interface equation	$\frac{n_1}{d_\text{o}} + \frac{n_2}{d_\text{i}} = \frac{n_2 - n_1}{R}$
The thin-lens equation	$\frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} = \frac{1}{f}$
The lens maker’s equation	$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$
The magnification m of an object	$m \equiv \frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}}$
Optical power	$P = \frac{1}{f}$
Optical power of thin, closely spaced lenses	$P_\text{total} = P_\text{lens1} + P_\text{lens2} + P_\text{lens3} + \cdots$
Angular magnification M of a simple magnifier	$M = \frac{\theta_\text{image}}{\theta_\text{object}}$
Angular magnification of an object a distance L from the eye for a convex lens of focal length	$M = \left(\frac{25\text{ cm}}{L}\right)\left(1 + \frac{L - \ell}{f}\right)$

f held a distance ℓ from the eye	
Range of angular magnification for a given lens for a person with a near point of 25 cm	$\frac{25\text{ cm}}{f} \leq M \leq 1 + \frac{25\text{ cm}}{f}$
Net magnification of compound microscope	$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{d_{\text{i}}^{\text{obj}}(f^{\text{eye}}+25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}}$

Conceptual Questions

Exercise:

Problem:

Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyze a microscope’s image?

Solution:

Microscopes create images of macroscopic size, so geometric optics applies.

Exercise:

Problem:

The image produced by the microscope in [link] cannot be projected. Could extra lenses or mirrors project it? Explain.

Exercise:

Problem:

If you want your microscope or telescope to project a real image onto a screen, how would you change the placement of the eyepiece relative to the objective?

Solution:

The eyepiece would be moved slightly farther from the objective so that the image formed by the objective falls just beyond the focal length of the eyepiece.

Problems

Exercise:

Problem:

A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the angular magnification of the eyepiece? (b) If there are two other objectives that can be used, having magnifications of 100 and 400, what other total magnifications are possible?

Exercise:

Problem:

(a) What magnification is produced by a 0.150 cm-focal length microscope objective that is 0.155 cm from the object being viewed? (b) What is the overall magnification if an $8\times$ eyepiece (one that produces an angular magnification of 8.00) is used?

Solution:

a. $\begin{aligned} \frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} &= \frac{1}{f} \Rightarrow d_\mathrm{i} = 4.65\ \mathrm{cm} \\ &\Rightarrow m = -30.0 \end{aligned}$;

b. $M_\mathrm{net} = -240$

Exercise:

Problem:

Where does an object need to be placed relative to a microscope for its 0.50 cm-focal length objective to produce a magnification of −400?

Exercise:

Problem:

An amoeba is 0.305 cm away from the 0.300 cm-focal length objective lens of a microscope. (a) Where is the image formed by the objective lens? (b) What is this image’s magnification? (c) An eyepiece with a 2.00-cm focal length is placed 20.0 cm from the objective. Where is the final image? (d) What angular magnification is produced by the eyepiece? (e) What is the overall magnification? (See [link].)

Solution:

a. $\frac{1}{d_{\text{o}}^{\text{obj}}} + \frac{1}{d_{\text{i}}^{\text{obj}}} = \frac{1}{f^{\text{obj}}} \Rightarrow d_{\text{i}}^{\text{obj}} = 18.3\text{ cm}$ behind the objective lens;
b. $m^{\text{obj}} = -60.0$;
c. $\begin{aligned} d_{\text{o}}^{\text{eye}} &= 1.70\text{ cm} \\ d_{\text{i}}^{\text{eye}} &= -11.3\text{ cm} \end{aligned}$
in front of the eyepiece; d. $M^{\text{eye}} = 13.5$;
e. $M_{\text{net}} = -810$

Exercise:

Problem:

Unreasonable Results Your friends show you an image through a microscope. They tell you that the microscope has an objective with a 0.500-cm focal length and an eyepiece with a 5.00-cm focal length. The resulting overall magnification is 250,000. Are these viable values for a microscope?

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

Exercise:

Problem:

What is the angular magnification of a telescope that has a 100 cm-focal length objective and a 2.50 cm-focal length eyepiece?

Solution:

$M = -40.0$

Exercise:

Problem:

Find the distance between the objective and eyepiece lenses in the telescope in the above problem needed to produce a final image very far from the observer, where vision is most relaxed. Note that a telescope is normally used to view very distant objects.

Exercise:

Problem:

A large reflecting telescope has an objective mirror with a 10.0-m radius of curvature. What angular magnification does it produce when a 3.00 m-focal length eyepiece is used?

Solution:

$f^{\text{obj}} = \frac{R}{2}, M = -1.67$

Exercise:

Problem:

A small telescope has a concave mirror with a 2.00-m radius of curvature for its objective. Its eyepiece is a 4.00 cm-focal length lens. (a) What is the telescope’s angular magnification? (b) What angle is subtended by a 25,000 km-diameter sunspot? (c) What is the angle of its telescopic image?

Exercise:

Problem:

A $7.5\ \times$ binocular produces an angular magnification of −7.50, acting like a telescope. (Mirrors are used to make the image upright.) If the binoculars have objective lenses with a 75.0-cm focal length, what is the focal length of the eyepiece lenses?

Solution:

$$M=-\frac{f^{\text{obj}}}{f^{\text{eye}}},f^{\text{eye}}=+10.0\,\text{cm}$$

Exercise:

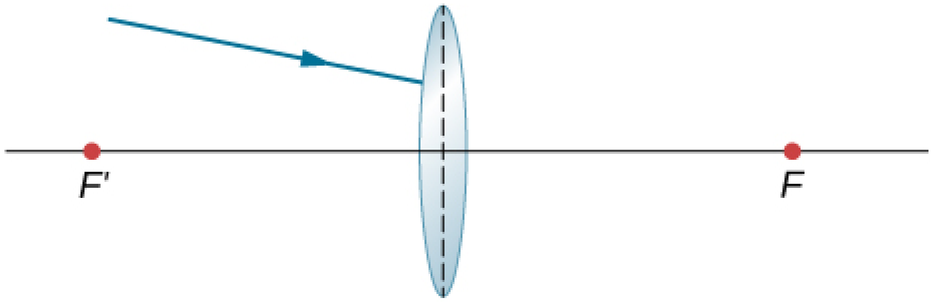
Problem:

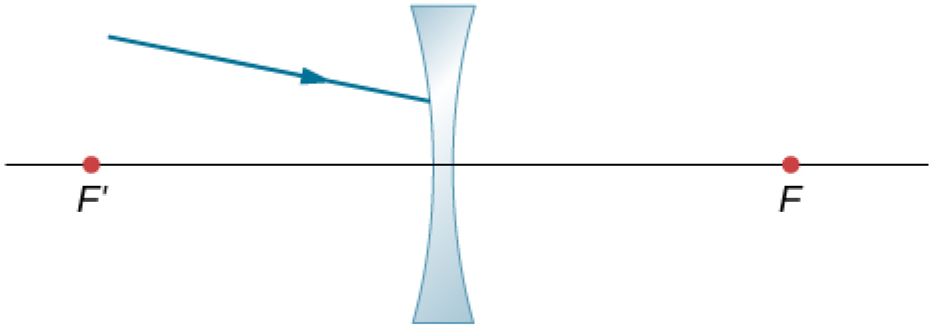
Construct Your Own Problem Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in part (a) of [link]. Construct a problem in which you calculate the location and size of the image produced. Among the things to be considered are the focal lengths of the lenses and their relative placements as well as the size and location of the object. Verify that the angular magnification is greater than one. That is, the angle subtended at the eye by the image is greater than the angle subtended by the object.

Exercise:

Problem:

Trace rays to find which way the given ray will emerge after refraction through the thin lens in the following figure. Assume thin-lens approximation. (*Hint*: Pick a point P on the given ray in each case. Treat that point as an object. Now, find its image Q. Use the rule: All rays on the other side of the lens will either go through Q or appear to be coming from Q.)





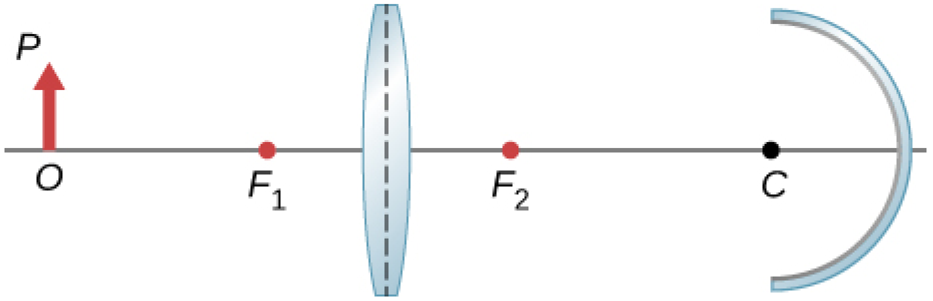
Solution:

Answers will vary.

Exercise:

Problem:

Copy and draw rays to find the final image in the following diagram. (*Hint*: Find the intermediate image through lens alone. Use the intermediate image as the object for the mirror and work with the mirror alone to find the final image.)



Exercise:

Problem:

A concave mirror of radius of curvature 10 cm is placed 30 cm from a thin convex lens of focal length 15 cm. Find the location and magnification of a small bulb sitting 50 cm from the lens by using the algebraic method.

Solution:

12 cm to the left of the mirror, $m = 3/5$

Exercise:

Problem:

An object of height 3 cm is placed at 25 cm in front of a converging lens of focal length 20 cm. Behind the lens there is a concave mirror of focal length 20 cm. The distance between the lens and the mirror is 5 cm. Find the location, orientation and size of the final image.

Exercise:

Problem:

An object of height 3 cm is placed at a distance of 25 cm in front of a converging lens of focal length 20 cm, to be referred to as the first lens. Behind the lens there is another converging lens of focal length 20 cm placed 10 cm from the first lens. There is a concave mirror of focal length 15 cm placed 50 cm from the second lens. Find the location, orientation, and size of the final image.

Solution:

27 cm in front of the mirror, $m = 0.6,\ h_i = 1.76$ cm, orientation upright

Exercise:

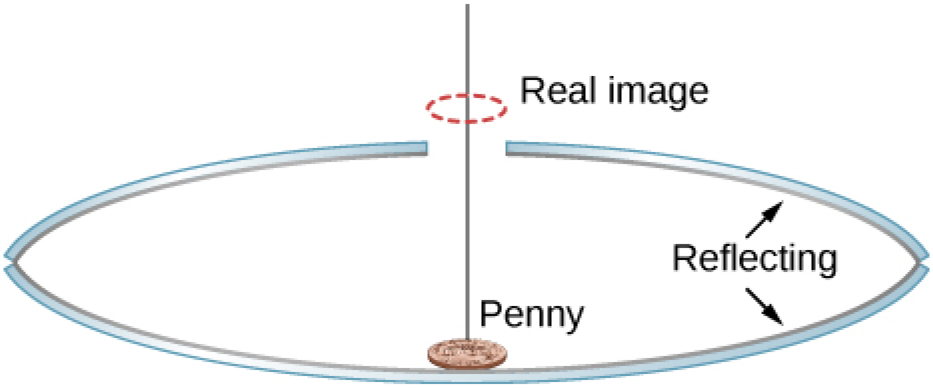
Problem:

An object of height 2 cm is placed at 50 cm in front of a diverging lens of focal length 40 cm. Behind the lens, there is a convex mirror of focal length 15 cm placed 30 cm from the converging lens. Find the location, orientation, and size of the final image.

Exercise:

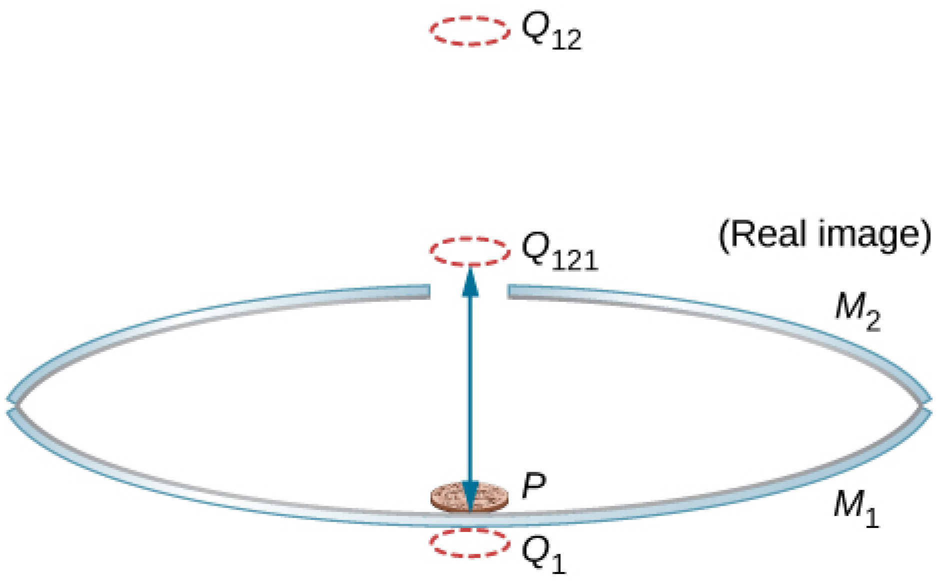
Problem:

Two concave mirrors are placed facing each other. One of them has a small hole in the middle. A penny is placed on the bottom mirror (see the following figure). When you look from the side, a real image of the penny is observed above the hole. Explain how that could happen.



Solution:

The following figure shows three successive images beginning with the image Q_1 in mirror M_1. Q_1 is the image in mirror M_1, whose image in mirror M_2 is Q_{12} whose image in mirror M_1 is the real image Q_{121}.



Exercise:

Problem:

A lamp of height 5 cm is placed 40 cm in front of a converging lens of focal length 20 cm. There is a plane mirror 15 cm behind the lens. Where would you find the image when you look in the mirror?

Exercise:

Problem:

Parallel rays from a faraway source strike a converging lens of focal length 20 cm at an angle of 15 degrees with the horizontal direction. Find the vertical position of the real image observed on a screen in the focal plane.

Solution:

5.4 cm from the axis

Exercise:

Problem:

Parallel rays from a faraway source strike a diverging lens of focal length 20 cm at an angle of 10 degrees with the horizontal direction. As you look through the lens, where in the vertical plane the image would appear?

Exercise:

Problem:

A light bulb is placed 10 cm from a plane mirror, which faces a convex mirror of radius of curvature 8 cm. The plane mirror is located at a distance of 30 cm from the vertex of the convex mirror. Find the location of two images in the convex mirror. Are there other images? If so, where are they located?

Solution:

Let the vertex of the concave mirror be the origin of the coordinate system. Image 1 is at −10/3 cm (−3.3 cm), image 2 is at −40/11 cm (−3.6 cm). These serve as objects for subsequent images, which are at −310/83 cm (−3.7 cm), −9340/2501 cm (−3.7 cm), −140,720/37,681 cm (−3.7 cm). All remaining images are at approximately −3.7 cm.

Exercise:

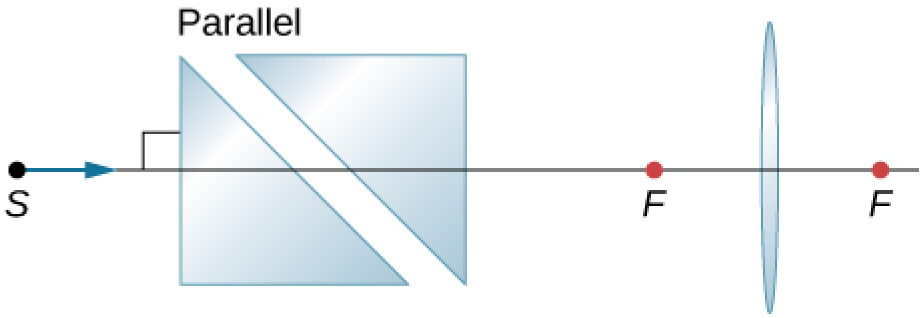
Problem:

A point source of light is 50 cm in front of a converging lens of focal length 30 cm. A concave mirror with a focal length of 20 cm is placed 25 cm behind the lens. Where does the final image form, and what are its orientation and magnification?

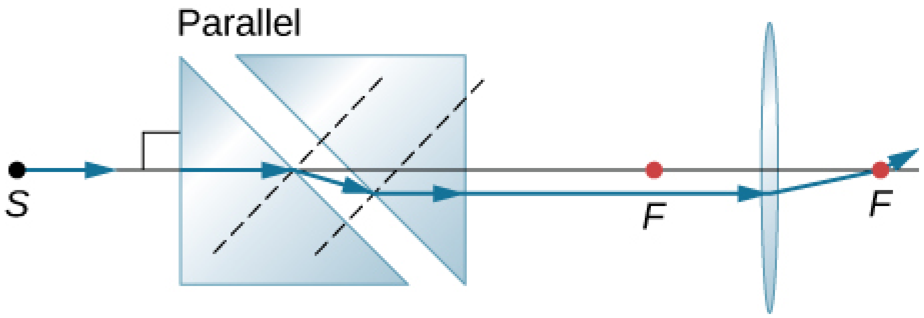
Exercise:

Problem:

Copy and trace to find how a horizontal ray from *S* comes out after the lens. Use $n_{\text{glass}} = 1.5$ for the prism material.


Parallel
S
F
F

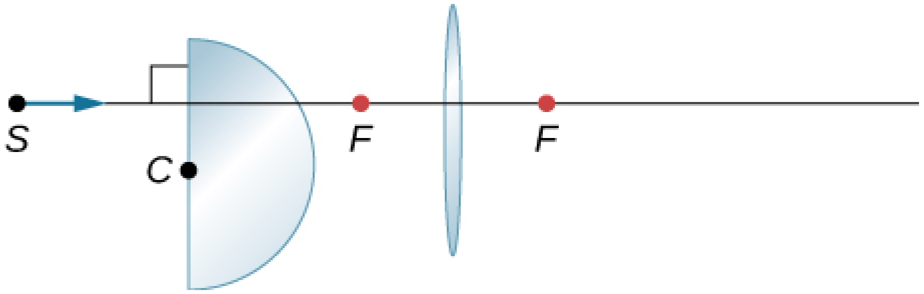
Solution:


Parallel
S
F
F

Exercise:

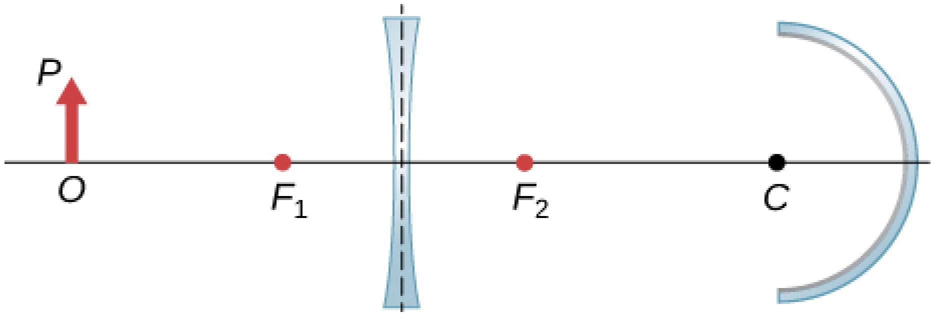
Problem:

Copy and trace how a horizontal ray from S comes out after the lens. Use $n = 1.55$ for the glass.

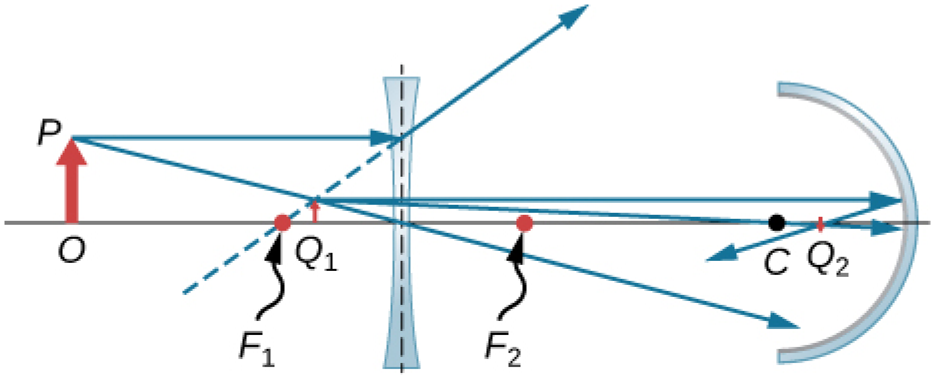


Exercise:

Problem: Copy and draw rays to figure out the final image.



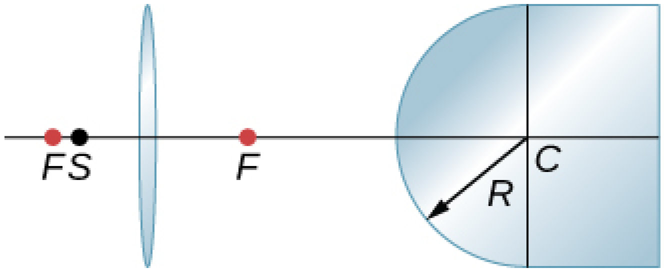
Solution:



Exercise:

Problem:

By ray tracing or by calculation, find the place inside the glass where rays from *S* converge as a result of refraction through the lens and the convex air-glass interface. Use a ruler to estimate the radius of curvature.



Exercise:

Problem:

A diverging lens has a focal length of 20 cm. What is the power of the lens in diopters?

Solution:

−5 D

Exercise:

Problem:

Two lenses of focal lengths of f_1 and f_2 are glued together with transparent material of negligible thickness. Show that the total power of the two lenses simply add.

Exercise:

Problem:

What will be the angular magnification of a convex lens with the focal length 2.5 cm?

Solution:

11

Exercise:

Problem:

What will be the formula for the angular magnification of a convex lens of focal length f if the eye is very close to the lens and the near point is located a distance D from the eye?

Additional Problems

Exercise:

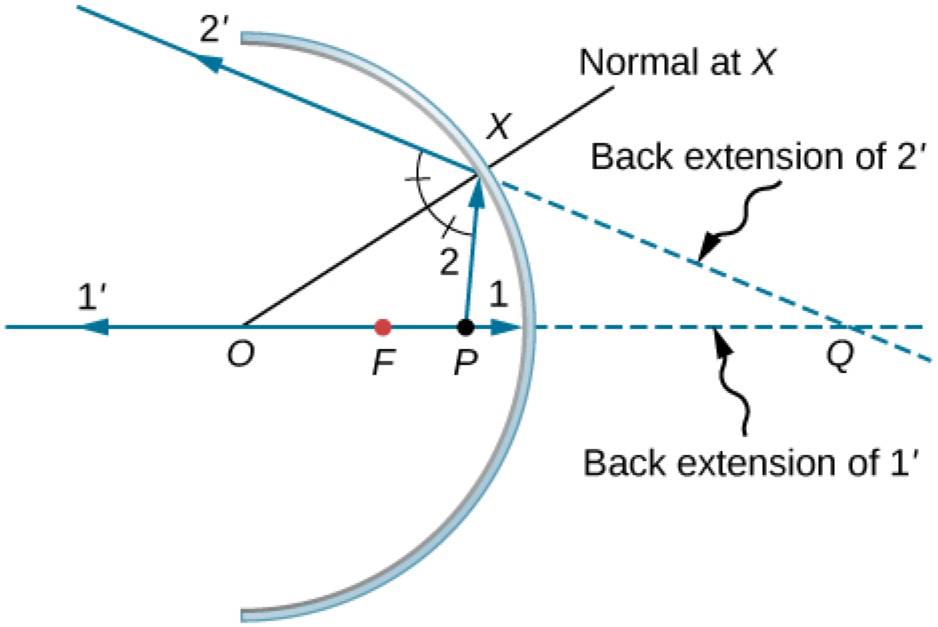
Problem:

Use a ruler and a protractor to draw rays to find images in the following cases.

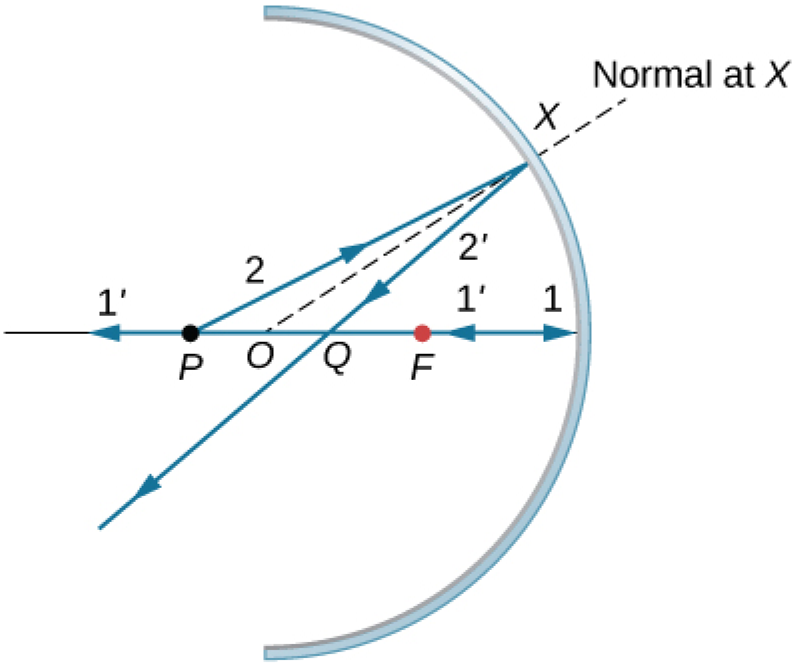
(a) A point object located on the axis of a concave mirror located at a point within the focal length from the vertex.
(b) A point object located on the axis of a concave mirror located at a point farther than the focal length from the vertex.
(c) A point object located on the axis of a convex mirror located at a point within the focal length from the vertex.
(d) A point object located on the axis of a convex mirror located at a point farther than the focal length from the vertex.
(e) Repeat (a)–(d) for a point object off the axis.

Solution:

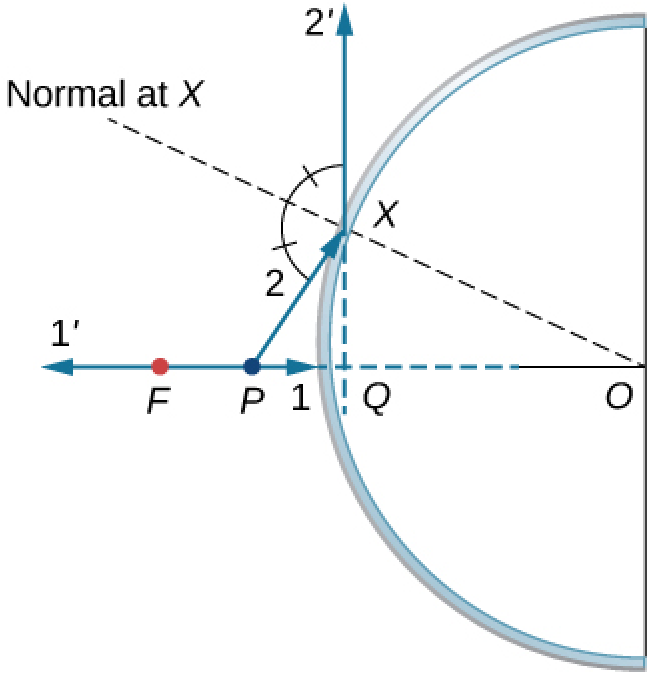
a.



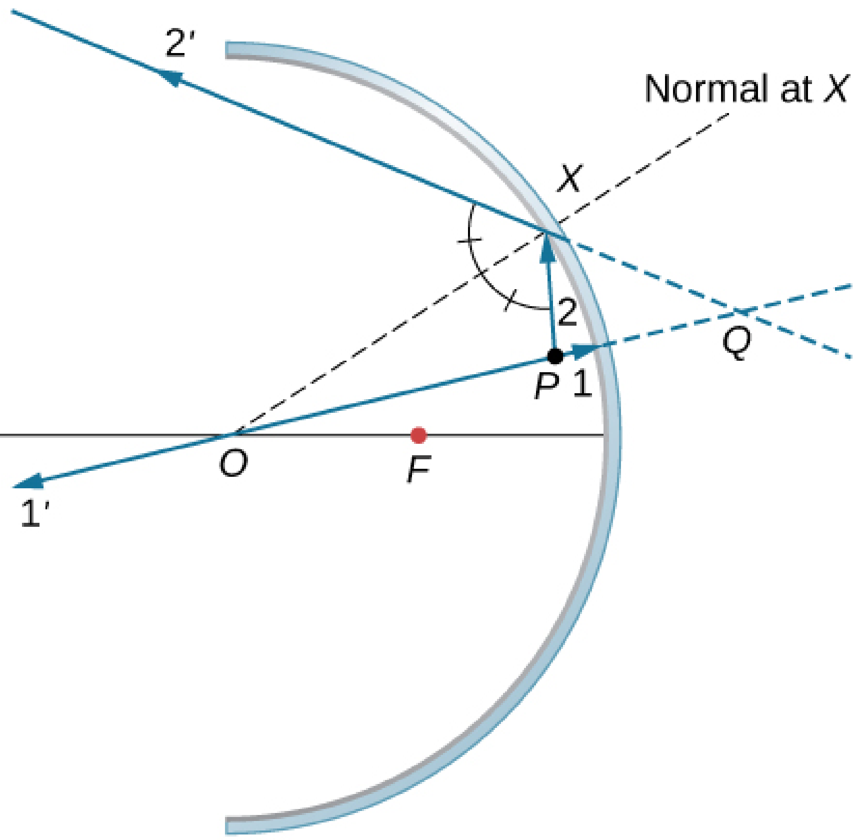
b.

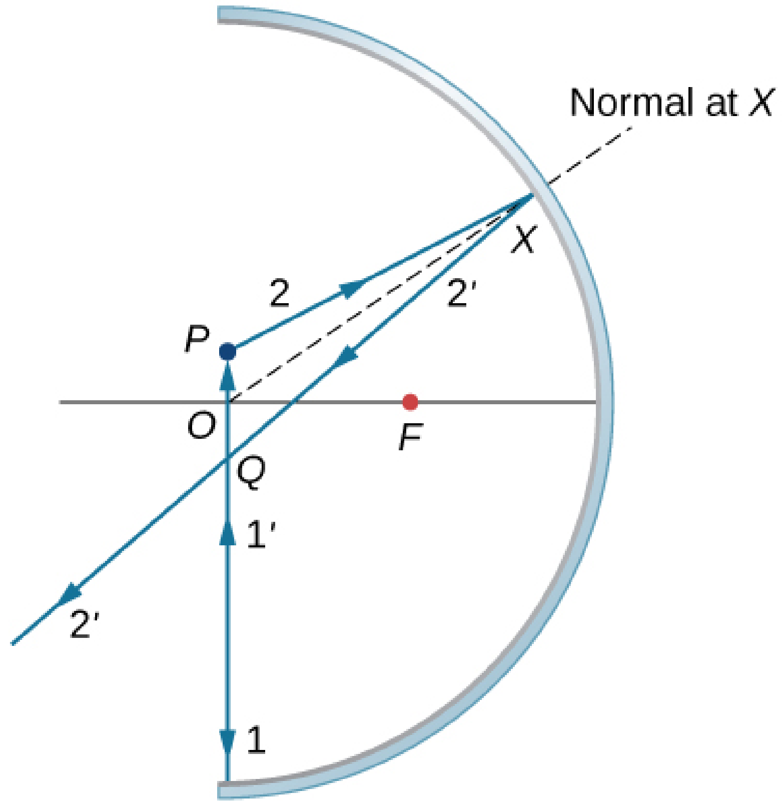

Normal at X
X
2
2′
1′
1′
1
P
O
Q
F

c.


2′
Normal at X
X
2
1′
F
P
1
Q
O

d. similar to the previous picture but with point *P* outside the focal length; e. Repeat (a)–(d) for a point object off the axis. For a point object placed off axis in front of a concave mirror corresponding to parts (a) and (b), the case for convex mirror left as exercises.


2′
Normal at X
X
2
Q
P
1
O
F
1′


Normal at X
X
2
2′
P
O
F
Q
1′
2′
1

Exercise:

Problem:

Where should a 3 cm tall object be placed in front of a concave mirror of radius 20 cm so that its image is real and 2 cm tall?

Exercise:

Problem:

A 3 cm tall object is placed 5 cm in front of a convex mirror of radius of curvature 20 cm. Where is the image formed? How tall is the image? What is the orientation of the image?

Solution:

$d_{\mathrm{i}} = -10/3\ \mathrm{cm}$, $h_{\mathrm{i}} = 2\ \mathrm{cm}$, upright

Exercise:

Problem:

You are looking for a mirror so that you can see a four-fold magnified virtual image of an object when the object is placed 5 cm from the vertex of the mirror. What kind of mirror you will need? What should be the radius of curvature of the mirror?

Exercise:

Problem: Derive the following equation for a convex mirror:

$$\frac{1}{VO} - \frac{1}{VI} = -\frac{1}{VF},$$

where *VO* is the distance to the object *O* from vertex *V*, *VI* the distance to the image *I* from *V*, and *VF* is the distance to the focal point *F* from *V*. (*Hint*: use two sets of similar triangles.)

Solution:

proof

Exercise:

Problem:

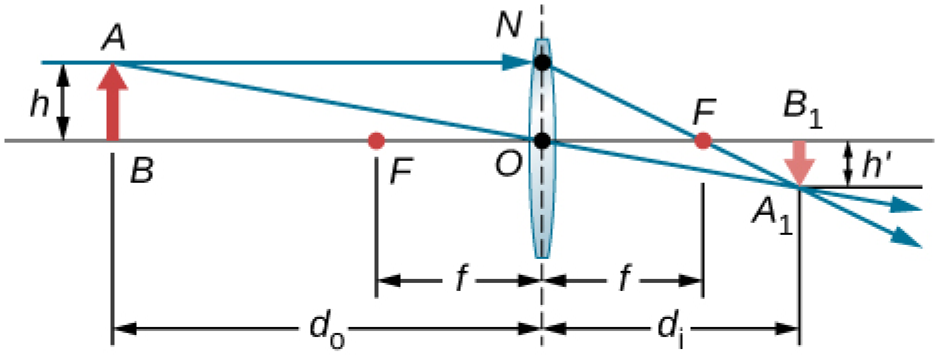
(a) Draw rays to form the image of a vertical object on the optical axis and farther than the focal point from a converging lens. (b) Use plane geometry in your figure and prove that the magnification *m* is given by $m=\frac{h_\text{i}}{h_\text{o}}=-\frac{d_\text{i}}{d_\text{o}}$.

Exercise:

Problem:

Use another ray-tracing diagram for the same situation as given in the previous problem to derive the thin-lens equation, $\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}=\frac{1}{f}$.

Solution:



Triangles BAO and B_1A_1O are similar triangles. Thus, $\frac{A_1B_1}{AB}=\frac{d_\text{i}}{d_\text{o}}$ .Triangles NOF and B_1A_1F are similar triangles. Thus, $\frac{NO}{f}=\frac{A_1B_1}{d_\text{i}-f}$. Noting that $NO=AB$ gives $\frac{AB}{f}=\frac{A_1B_1}{d_\text{i}-f}$ or $\frac{AB}{A_1B_1}=\frac{f}{d_\text{i}-f}$. Inverting this gives $\frac{A_1B_1}{AB}=\frac{d_\text{i}-f}{f}$. Equating the two expressions for the ratio $\frac{A_1B_1}{AB}$ gives $\frac{d_\text{i}}{d_\text{o}}=\frac{d_\text{i}-f}{f}$. Dividing through by d_i gives $\frac{1}{d_\text{o}}=\frac{1}{f}-\frac{1}{d_\text{i}}$ or $\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}=\frac{1}{f}$.

Exercise:

Problem:

You photograph a 2.0-m-tall person with a camera that has a 5.0 cm-focal length lens. The image on the film must be no more than 2.0 cm high. (a) What is the closest distance the person can stand to the lens? (b) For this distance, what should be the distance from the lens to the film?

Exercise:

Problem:

Find the focal length of a thin plano-convex lens. The front surface of this lens is flat, and the rear surface has a radius of curvature of $R_2 = -35\text{ cm}$. Assume that the index of refraction of the lens is 1.5.

Solution:

70 cm

Exercise:

Problem:

Find the focal length of a meniscus lens with $R_1 = 20\text{ cm}$ and $R_2 = 15\text{ cm}$. Assume that the index of refraction of the lens is 1.5.

Exercise:

Problem:

A nearsighted man cannot see objects clearly beyond 20 cm from his eyes. How close must he stand to a mirror in order to see what he is doing when he shaves?

Solution:

The plane mirror has an infinite focal point, so that $d_\text{i} = -d_\text{o}$. The total apparent distance of the man in the mirror will be his actual distance, plus the apparent image distance, or $d_\text{o} + (-d_\text{i}) = 2d_\text{o}$. If this distance must be less than 20 cm, he should stand at $d_\text{o} = 10\text{ cm}$.

Exercise:

Problem:

A mother sees that her child’s contact lens prescription is 0.750 D. What is the child’s near point?

Exercise:

Problem:

Repeat the previous problem for glasses that are 2.20 cm from the eyes.

Solution:

Here we want $d_{\text{o}} = 25\text{ cm} - 2.20\text{ cm} = 0.228\text{ m}$. If $x = \text{near point}$, $d_{\text{i}} = -(x - 0.0220\text{ m})$. Thus, $P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{0.228\text{ m}} + \frac{1}{x - 0.0220\text{ m}}$. Using $P = 0.75\text{ D}$ gives $x = 0.253\text{ m}$, so the near point is 25.3 cm.

Exercise:

Problem:

The contact-lens prescription for a nearsighted person is −4.00 D and the person has a far point of 22.5 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Exercise:

Problem:

Unreasonable Results A boy has a near point of 50 cm and a far point of 500 cm. Will a −4.00 D lens correct his far point to infinity?

Solution:

Assuming a lens at 2.00 cm from the boy’s eye, the image distance must be $d_{\text{i}} = -(500\text{ cm} - 2.00\text{ cm}) = -498\text{ cm}$. For an infinite-distance object, the required power is $P = \frac{1}{d_\text{i}} = -0.200\,\text{D}$. Therefore, the $-4.00\,\text{D}$ lens will correct the nearsightedness.

Exercise:

Problem:

Find the angular magnification of an image by a magnifying glass of $f = 5.0\,\text{cm}$ if the object is placed $d_\text{o} = 4.0\,\text{cm}$ from the lens and the lens is close to the eye.

Exercise:

Problem:

Let objective and eyepiece of a compound microscope have focal lengths of 2.5 cm and 10 cm, respectively and be separated by 12 cm. A $70\text{-}\mu\text{m}$ object is placed 6.0 cm from the objective. How large is the virtual image formed by the objective-eyepiece system?

Solution:

$87\,\mu\text{m}$

Exercise:

Problem:

Draw rays to scale to locate the image at the retina if the eye lens has a focal length 2.5 cm and the near point is 24 cm. (*Hint*: Place an object at the near point.)

Exercise:

Problem:

The objective and the eyepiece of a microscope have the focal lengths 3 cm and 10 cm respectively. Decide about the distance between the objective and the eyepiece if we need a $10\,\times$ magnification from the objective/eyepiece compound system.

Solution:

Use, $M_{\text{net}} = -\frac{d_{\text{i}}^{\text{obj}}(f^{\text{eye}}+25\text{ cm})}{f^{\text{obj}}f^{\text{eye}}}$. The image distance for the objective is $d_{\text{i}}^{\text{obj}} = -\frac{M_{\text{net}}f^{\text{obj}}f^{\text{eye}}}{f^{\text{eye}}+25\text{ cm}}$. Using $f^{\text{obj}} = 3.0\text{ cm}, f^{\text{eye}} = 10\text{ cm}, \text{and } M = -10 \text{ gives } d_{\text{i}}^{\text{obj}} = 8.6\text{ cm}.$ We want this image to be at the focal point of the eyepiece so that the eyepiece forms an image at infinity for comfortable viewing. Thus, the distance *d* between the lenses should be $d = f^{\text{eye}} + d_{\text{i}}^{\text{obj}} = 10\text{ cm} + 8.6\text{ cm} = 19\text{ cm}.$

Exercise:

Problem:

A far-sighted person has a near point of 100 cm. How far in front or behind the retina does the image of an object placed 25 cm from the eye form? Use the cornea to retina distance of 2.5 cm.

Exercise:

Problem:

A near-sighted person has afar point of 80 cm. (a) What kind of corrective lens the person will need if the lens is to be placed 1.5 cm from the eye? (b) What would be the power of the contact lens needed? Assume distance to contact lens from the eye to be zero.

Solution:

a. focal length of the corrective lens $f_{\text{c}} = -80$ cm; b. −1.25 D

Exercise:

Problem:

In a reflecting telescope the objective is a concave mirror of radius of curvature 2 m and an eyepiece is a convex lens of focal length 5 cm. Find the apparent size of a 25-m tree at a distance of 10 km that you would perceive when looking through the telescope.

Exercise:

Problem:

Two stars that are 10^{9}km apart are viewed by a telescope and found to be separated by an angle of 10^{-5} radians. If the eyepiece of the telescope has a focal length of 1.5 cm and the objective has a focal length of 3 meters, how far away are the stars from the observer?

Solution:

$2\;\times\;10^{16}\;\text{km}$

Exercise:

Problem:

What is the angular size of the Moon if viewed from a binocular that has a focal length of 1.2 cm for the eyepiece and a focal length of 8 cm for the objective? Use the radius of the moon $1.74\;\times\;10^{6}\text{m}$ and the distance of the moon from the observer to be $3.8\;\times\;10^{8}\text{m}$.

Exercise:

Problem:

An unknown planet at a distance of 10^{12}m from Earth is observed by a telescope that has a focal length of the eyepiece of 1 cm and a focal length of the objective of 1 m. If the far away planet is seen to subtend an angle of 10^{-5} radian at the eyepiece, what is the size of the planet?

Solution:

$10^{5}\;\text{m}$

Glossary

Cassegrain design

arrangement of an objective and eyepiece such that the light-gathering concave mirror has a hole in the middle, and light then is incident on an eyepiece lens

compound microscope
microscope constructed from two convex lenses, the first serving as the eyepiece and the second serving as the objective lens

eyepiece
lens or combination of lenses in an optical instrument nearest to the eye of the observer

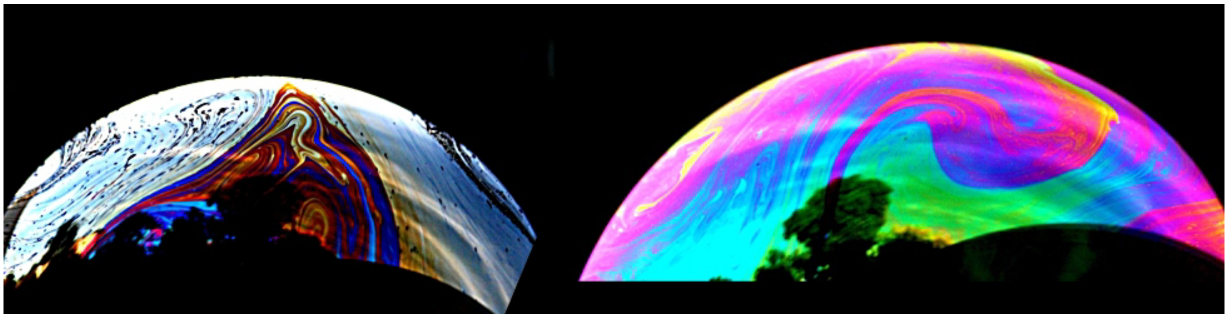
net magnification
(M_{net}) of the compound microscope is the product of the linear magnification of the objective and the angular magnification of the eyepiece

Newtonian design
arrangement of an objective and eyepiece such that the focused light from the concave mirror was reflected to one side of the tube into an eyepiece

objective
lens nearest to the object being examined.

Introduction
class="introduction"

Soap bubbles are blown from clear fluid into very thin films. The colors we see are not due to any pigmentation but are the result of light interference, which enhances specific wavelengths for a given thickness of the film.



The most certain indication of a wave is interference. This wave characteristic is most prominent when the wave interacts with an object that is not large compared with the wavelength. Interference is observed for water waves, sound waves, light waves, and, in fact, all types of waves.

If you have ever looked at the reds, blues, and greens in a sunlit soap bubble and wondered how straw-colored soapy water could produce them, you have hit upon one of the many phenomena that can only be explained by the wave character of light (see [link]). The same is true for the colors seen in an oil slick or in the light reflected from a DVD disc. These and other interesting phenomena cannot be explained fully by geometric optics. In these cases, light interacts with objects and exhibits wave characteristics. The branch of optics that considers the behavior of light when it exhibits wave characteristics is called wave optics (sometimes called physical optics). It is the topic of this chapter.

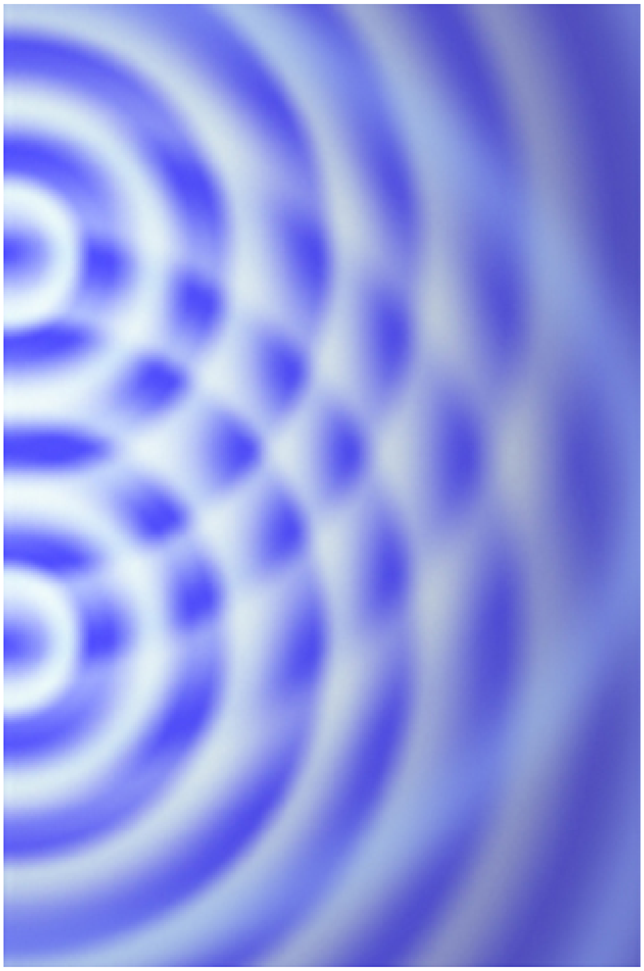
Young's Double-Slit Interference

By the end of this section, you will be able to:

- Explain the phenomenon of interference
- Define constructive and destructive interference for a double slit

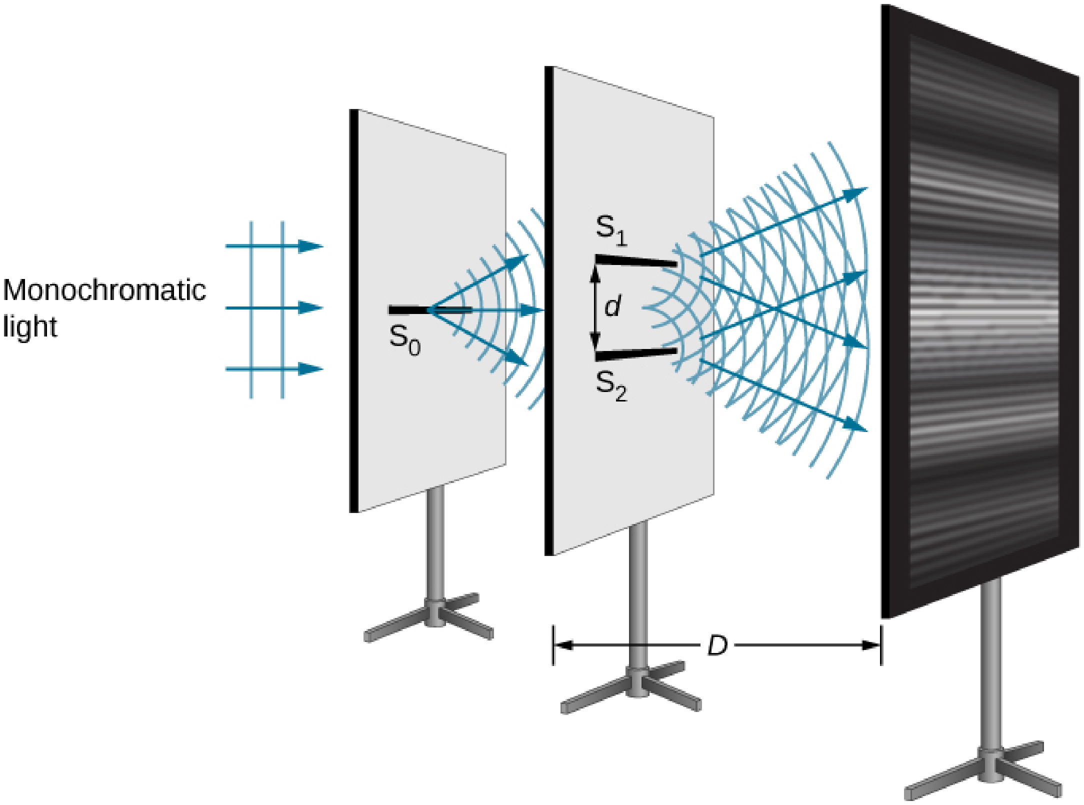
The Dutch physicist Christiaan Huygens (1629–1695) thought that light was a wave, but Isaac Newton did not. Newton thought that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton’s tremendous reputation, his view generally prevailed; the fact that Huygens’s principle worked was not considered direct evidence proving that light is a wave. The acceptance of the wave character of light came many years later in 1801, when the English physicist and physician Thomas Young (1773–1829) demonstrated optical interference with his now-classic double-slit experiment.

If there were not one but two sources of waves, the waves could be made to interfere, as in the case of waves on water ([link]). If light is an electromagnetic wave, it must therefore exhibit interference effects under appropriate circumstances. In Young’s experiment, sunlight was passed through a pinhole on a board. The emerging beam fell on two pinholes on a second board. The light emanating from the two pinholes then fell on a screen where a pattern of bright and dark spots was observed. This pattern, called fringes, can only be explained through interference, a wave phenomenon.



Photograph of an interference pattern produced by circular water waves in a ripple tank. Two thin plungers are vibrated up and down in phase at the surface of the water. Circular water waves are produced by and emanate from each plunger. The points where the water is calm (corresponding to destructive interference) are clearly visible.

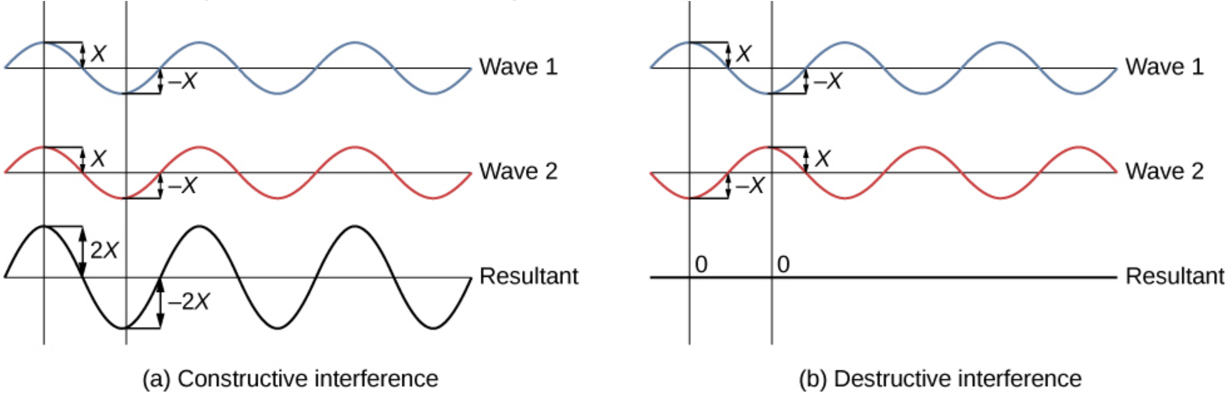
We can analyze double-slit interference with the help of [link], which depicts an apparatus analogous to Young’s. Light from a monochromatic source falls on a slit S_0. The light emanating from S_0 is incident on two other slits S_1 and S_2 that are equidistant from S_0. A pattern of *interference fringes* on the screen is then produced by the light emanating from S_1 and S_2. All slits are assumed to be so narrow that they can be considered secondary point sources for Huygens' wavelets (The Nature of Light). Slits S_1 and S_2 are a distance d apart ($d \leq 1\ \mathrm{mm}$), and the distance between the screen and the slits is $D(\approx 1\ \mathrm{m})$, which is much greater than d.



The double-slit interference experiment using monochromatic light and narrow slits. Fringes produced by interfering Huygens wavelets from slits S_1 and S_2 are observed on the screen.

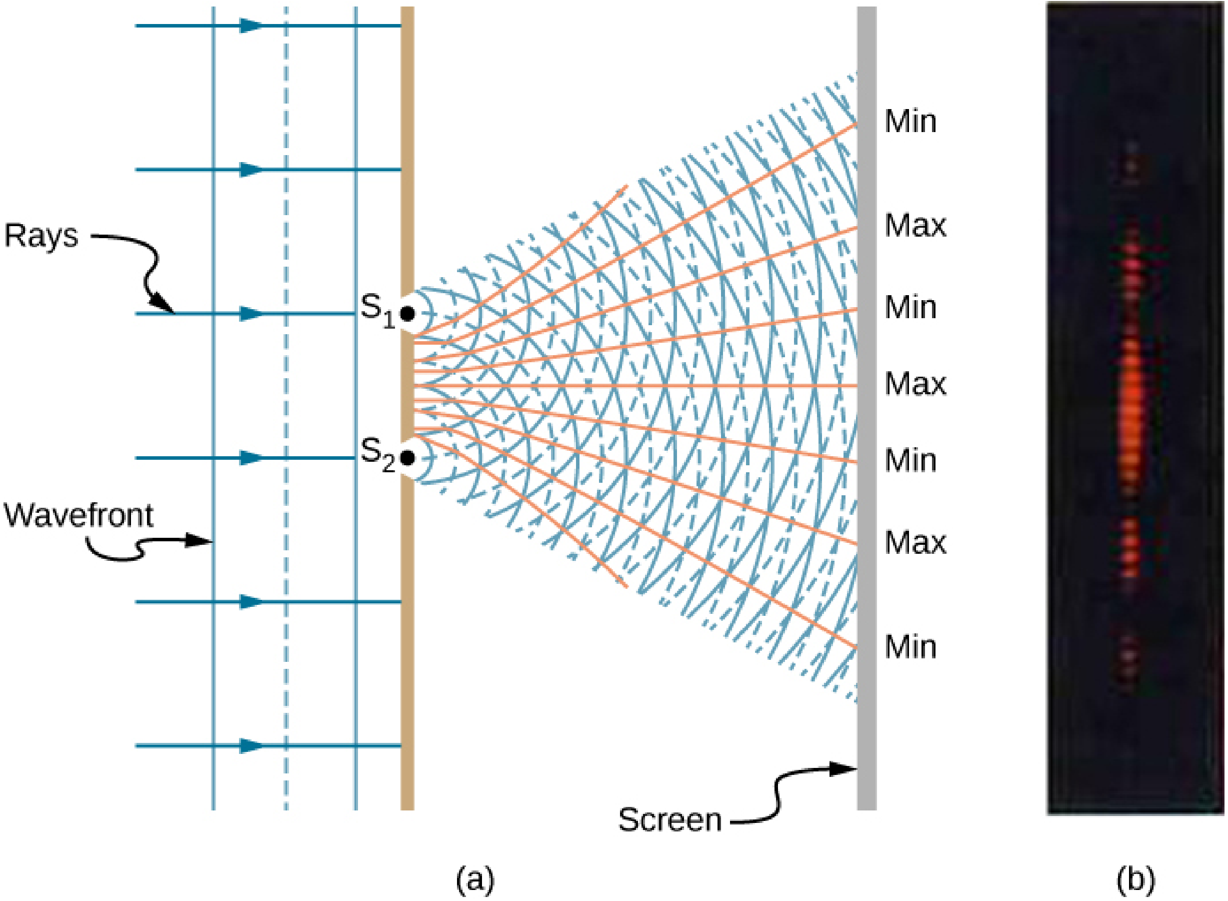
Since S_0 is assumed to be a point source of monochromatic light, the secondary Huygens wavelets leaving S_1 and S_2 always maintain a constant phase difference (zero in this case because S_1 and S_2 are equidistant from S_0) and have the same frequency. The sources S_1 and S_2 are then said to be coherent. By **coherent waves**, we mean the waves are in phase or have a definite phase relationship. The term **incoherent** means the waves have random phase relationships, which would be the case if S_1 and S_2 were illuminated by two independent light sources, rather than a single source S_0. Two independent light sources (which may be two separate areas within the same lamp or the Sun) would generally not emit their light in unison, that is, not coherently. Also, because S_1 and S_2 are the same distance from S_0, the amplitudes of the two Huygens wavelets are equal.

Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. In the following discussion, we illustrate the double-slit experiment with **monochromatic** light (single λ) to clarify the effect. [link] shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.



The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

When light passes through narrow slits, the slits act as sources of coherent waves and light spreads out as semicircular waves, as shown in [link](a). Pure *constructive interference* occurs where the waves are crest to crest or trough to trough. Pure *destructive interference* occurs where they are crest to trough. The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in [link]. Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.



Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) When light that has passed through double slits falls on a screen, we see a pattern such as this.

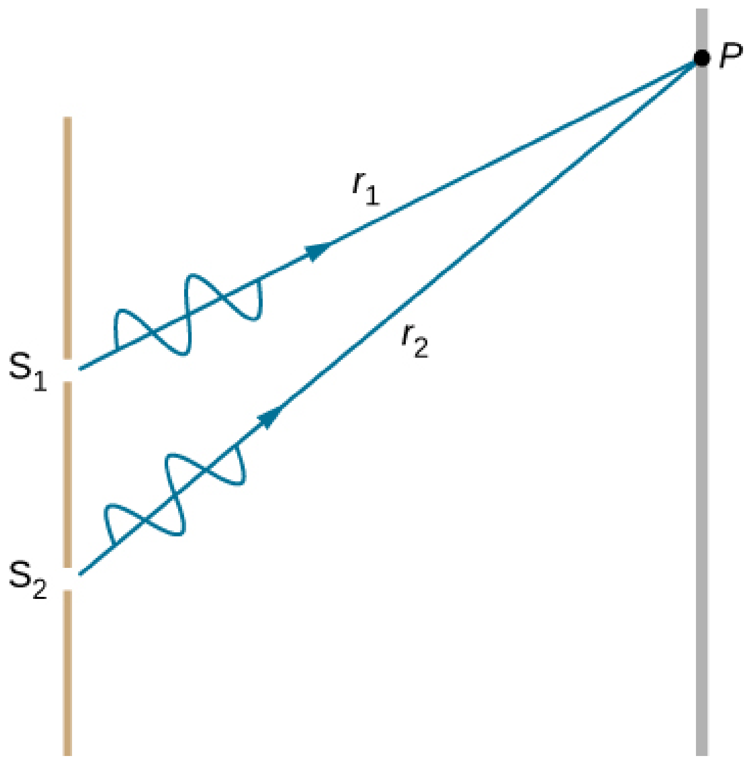
To understand the double-slit interference pattern, consider how two waves travel from the slits to the screen ([link]). Each slit is a different distance from a given point on the screen. Thus, different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively. If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively. More generally, if the path length difference Δl between the two waves is any half-integral number of wavelengths [(1 / 2)λ, (3 / 2)λ, (5 / 2)λ, etc.], then destructive interference occurs. Similarly, if the path length difference is any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs. These conditions can be expressed as equations:

Equation:

$$\Delta l = m\lambda, \quad \text{for } m = 0, \pm 1, \pm 2, \pm 3 \ \ldots \ \text{(constructive interference)}$$

Equation:

$$\Delta l = (m + \frac{1}{2})\lambda, \quad \text{for } m = 0, \pm 1, \pm 2, \pm 3 \ \ldots \ \text{(destructive interference)}$$



Waves follow different paths from the slits to a common point *P* on a screen. Destructive interference occurs where one path is a half wavelength longer than the other—the waves start in phase but arrive out of phase. Constructive interference occurs where one path is a whole wavelength longer than the other —the waves start out and arrive in phase.

Summary

- Young’s double-slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.

Conceptual Questions

Exercise:

Problem:

Young’s double-slit experiment breaks a single light beam into two sources. Would the same pattern be obtained for two independent sources of light, such as the headlights of a distant car? Explain.

Solution:

No. Two independent light sources do not have coherent phase.

Exercise:

Problem:

Is it possible to create a experimental setup in which there is only destructive interference? Explain.

Exercise:

Problem:

Why won’t two small sodium lamps, held close together, produce an interference pattern on a distant screen? What if the sodium lamps were replaced by two laser pointers held close together?

Solution:

Because both the sodium lamps are not coherent pairs of light sources. Two lasers operating independently are also not coherent so no interference pattern results.

Glossary

coherent waves
    waves are in phase or have a definite phase relationship

incoherent
    waves have random phase relationships

monochromatic
    light composed of one wavelength only

Mathematics of Interference

By the end of this section, you will be able to:

- Determine the angles for bright and dark fringes for double slit interference
- Calculate the positions of bright fringes on a screen

[link](a) shows how to determine the path length difference Δl for waves traveling from two slits to a common point on a screen. If the screen is a large distance away compared with the distance between the slits, then the angle θ between the path and a line from the slits to the screen [part (b)] is nearly the same for each path. In other words, r_1 and r_2 are essentially parallel. The lengths of r_1 and r_2 differ by Δl, as indicated by the two dashed lines in the figure. Simple trigonometry shows

Equation:

$$\Delta l = d \sin \theta$$

where *d* is the distance between the slits. Combining this result with [link], we obtain constructive interference for a double slit when the path length difference is an integral multiple of the wavelength, or

Equation:

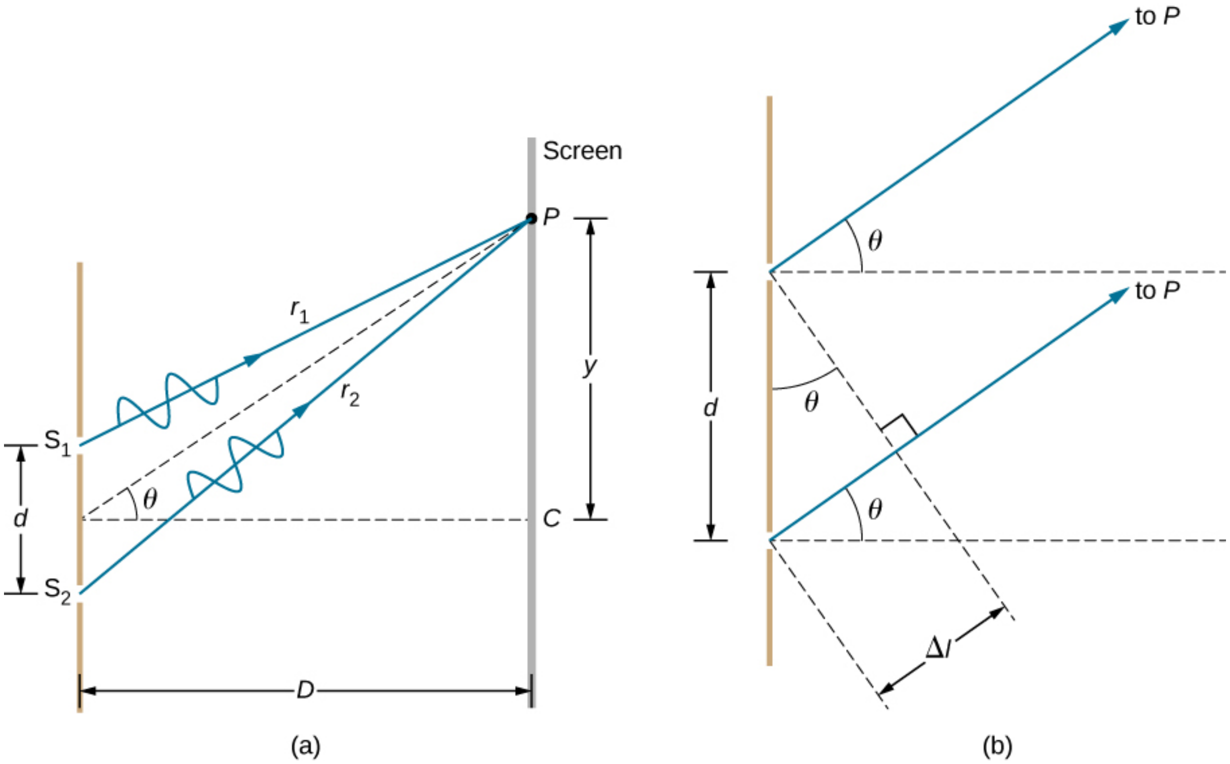
$$d \sin \theta = m\lambda, \text{ for } m = 0, \pm 1, \pm 2, \pm 3, \ldots \text{ (constructive interference).}$$

Similarly, to obtain destructive interference for a double slit, the path length difference must be a half-integral multiple of the wavelength, or

Equation:

$$d \sin \theta = (m + \frac{1}{2})\lambda, \text{ for } m = 0, \pm 1, \pm 2, \pm 3, \ldots \text{ (destructive interference)}$$

where λ is the wavelength of the light, *d* is the distance between slits, and θ is the angle from the original direction of the beam as discussed above. We call *m* the **order** of the interference. For example, $m = 4$ is fourth-order interference.



(a) To reach P, the light waves from S_1 and S_2 must travel different distances. (b) The path difference between the two rays is Δl.

The equations for double-slit interference imply that a series of bright and dark lines are formed. For vertical slits, the light spreads out horizontally on either side of the incident beam into a pattern called interference **fringes** ([link]). The closer the slits are, the more the bright fringes spread apart. We can see this by examining the equation

$d \sin \theta = m\lambda, \text{ for } m = 0, \pm 1, \pm 2, \pm 3 \ldots$. For fixed λ and m, the smaller d is, the larger θ must be, since $\sin \theta = m\lambda/d$. This is consistent with our contention that wave effects are most noticeable when the object the wave encounters (here, slits a distance d apart) is small. Small d gives large θ, hence, a large effect.

Referring back to part (a) of the figure, θ is typically small enough that $\sin \theta \approx \tan \theta \approx y_m/D$, where y_m is the distance from the central maximum to the mth bright fringe and D is the distance between the slit and the screen. [link] may then be written as

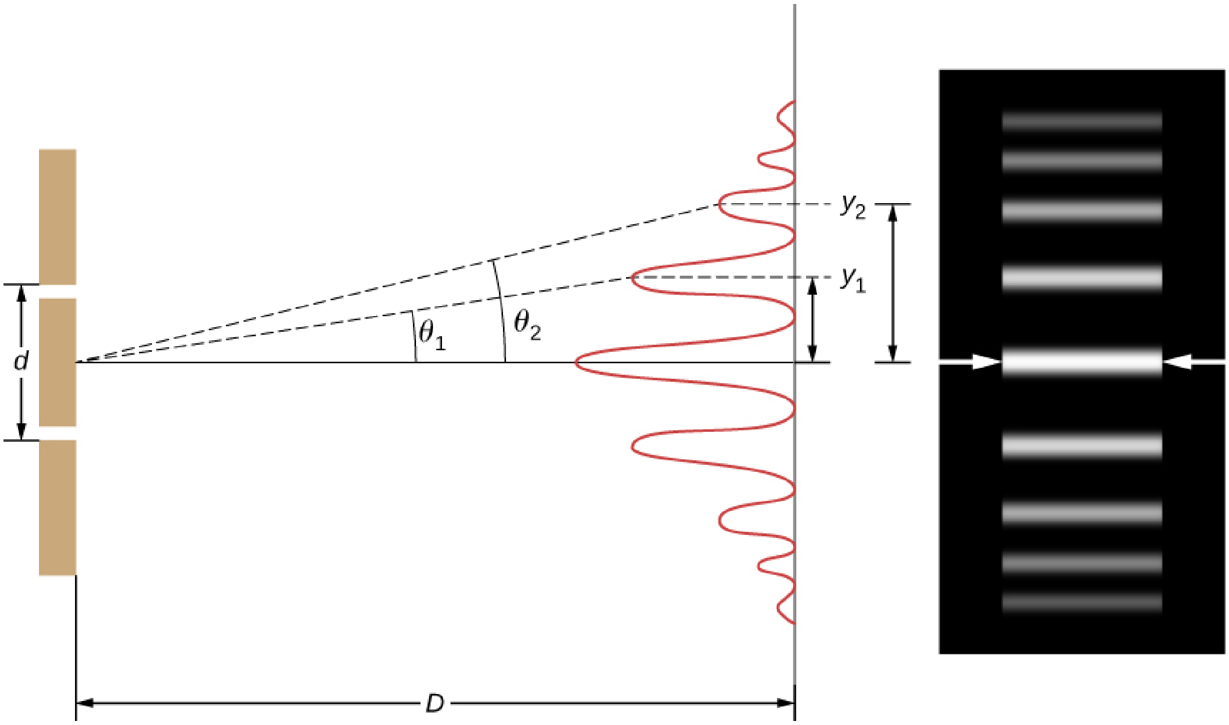
Equation:

$$d\frac{y_m}{D} = m\lambda$$

or

Equation:

$$y_m = \frac{m\lambda D}{d}.$$



The interference pattern for a double slit has an intensity that falls off with angle. The image shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example:
Finding a Wavelength from an Interference Pattern

Suppose you pass light from a He-Ne laser through two slits separated by 0.0100 mm and find that the third bright line on a screen is formed at an angle of 10.95° relative to the incident beam. What is the wavelength of the light?

Strategy

The phenomenon is two-slit interference as illustrated in [link] and the third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d = 0.0100\text{ mm}$ and $\theta = 10.95^\circ$. The wavelength can thus be found using the equation $d\sin\theta = m\lambda$ for constructive interference.

Solution

Solving $d\sin\theta = m\lambda$ for the wavelength λ gives

Equation:

$$\lambda = \frac{d\sin\theta}{m}.$$

Substituting known values yields

Equation:

$$\lambda = \frac{(0.0100\text{ mm})(\sin 10.95^\circ)}{3} = 6.33 \times 10^{-4}\text{ mm} = 633\text{ nm}.$$

Significance

To three digits, this is the wavelength of light emitted by the common He-Ne laser. Not by coincidence, this red color is similar to that emitted by neon lights. More important, however, is the fact that interference patterns can be used to measure wavelength. Young did this for visible wavelengths. This analytical techinque is still widely used to measure electromagnetic spectra. For a given order, the angle for constructive interference increases with λ, so that spectra (measurements of intensity versus wavelength) can be obtained.

Example:

Calculating the Highest Order Possible

Interference patterns do not have an infinite number of lines, since there is a limit to how big *m* can be. What is the highest-order constructive interference possible with the system described in the preceding example?

Strategy

The equation $d \sin \theta = m\lambda$ (for $m = 0, \pm 1, \pm 2, \pm 3...$) describes constructive interference from two slits. For fixed values of d and λ, the larger m is, the larger $\sin \theta$ is. However, the maximum value that $\sin \theta$ can have is 1, for an angle of 90°. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find what value of m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d \sin \theta = m\lambda$ for m gives

Equation:

$$m = \frac{d \sin \theta}{\lambda}.$$

Taking $\sin \theta = 1$ and substituting the values of d and λ from the preceding example gives

Equation:

$$m = \frac{(0.0100 \text{ mm})(1)}{633 \text{ nm}} \approx 15.8.$$

Therefore, the largest integer m can be is 15, or $m = 15$.

Significance

The number of fringes depends on the wavelength and slit separation. The number of fringes is very large for large slit separations. However, recall (see The Propagation of Light and the introduction for this chapter) that wave interference is only prominent when the wave interacts with objects that are not large compared to the wavelength. Therefore, if the slit separation and the sizes of the slits become much greater than the wavelength, the intensity pattern of light on the screen changes, so there are simply two bright lines cast by the slits, as expected, when light behaves like rays. We also note that the fringes get fainter farther away from the center. Consequently, not all 15 fringes may be observable.

Note:

Exercise:

Problem:

Check Your Understanding In the system used in the preceding examples, at what angles are the first and the second bright fringes formed?

Solution:

3.63° and 7.27°, respectively

Summary

- In double-slit diffraction, constructive interference occurs when $d\sin\theta = m\lambda$ (for $m = 0, \pm1, \pm2, \pm3\ldots$), where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.
- Destructive interference occurs when $d\sin\theta = (m + \frac{1}{2})\lambda$ for $m\ =\ 0, \pm1, \pm2, \pm3, \ldots$.

Conceptual Questions

Exercise:

Problem:

Suppose you use the same double slit to perform Young’s double-slit experiment in air and then repeat the experiment in water. Do the angles to the same parts of the interference pattern get larger or smaller? Does the color of the light change? Explain.

Exercise:

Problem:

Why is monochromatic light used in the double slit experiment? What would happen if white light were used?

Solution:

Monochromatic sources produce fringes at angles according to $d \sin \theta = m\lambda$. With white light, each constituent wavelength will produce fringes at its own set of angles, blending into the fringes of adjacent wavelengths. This results in rainbow patterns.

Problems

Exercise:

Problem:

At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?

Exercise:

Problem:

Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.

Solution:

0.997°

Exercise:

Problem:

What is the separation between two slits for which 610-nm orange light has its first maximum at an angle of 30.0°?

Exercise:

Problem:

Find the distance between two slits that produces the first minimum for 410-nm violet light at an angle of 45.0°.

Solution:

$0.290\ \mu\text{m}$

Exercise:

Problem:

Calculate the wavelength of light that has its third minimum at an angle of 30.0° when falling on double slits separated by $3.00\ \mu\text{m}$. Explicitly show how you follow the steps from the Problem-Solving Strategy: Wave Optics, located at the end of the chapter.

Exercise:

Problem:

What is the wavelength of light falling on double slits separated by $2.00\ \mu\text{m}$ if the third-order maximum is at an angle of 60.0°?

Solution:

$5.77\ \times\ 10^{-7}\,\text{m} = 577\,\text{nm}$

Exercise:

Problem:

At what angle is the fourth-order maximum for the situation in the preceding problem?

Exercise:

Problem:

What is the highest-order maximum for 400-nm light falling on double slits separated by $25.0\ \mu\text{m}$?

Solution:

62.5; since *m* must be an integer, the highest order is then $m = 62$.

Exercise:

Problem:

Find the largest wavelength of light falling on double slits separated by $1.20\ \mu\text{m}$ for which there is a first-order maximum. Is this in the visible part of the spectrum?

Exercise:

Problem:

What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?

Solution:

$1.44\ \mu\text{m}$

Exercise:

Problem:

(a) What is the smallest separation between two slits that will produce a second-order maximum for any visible light? (b) For all visible light?

Exercise:

Problem:

(a) If the first-order maximum for monochromatic light falling on a double slit is at an angle of 10.0°, at what angle is the second-order maximum? (b) What is the angle of the first minimum? (c) What is the highest-order maximum possible here?

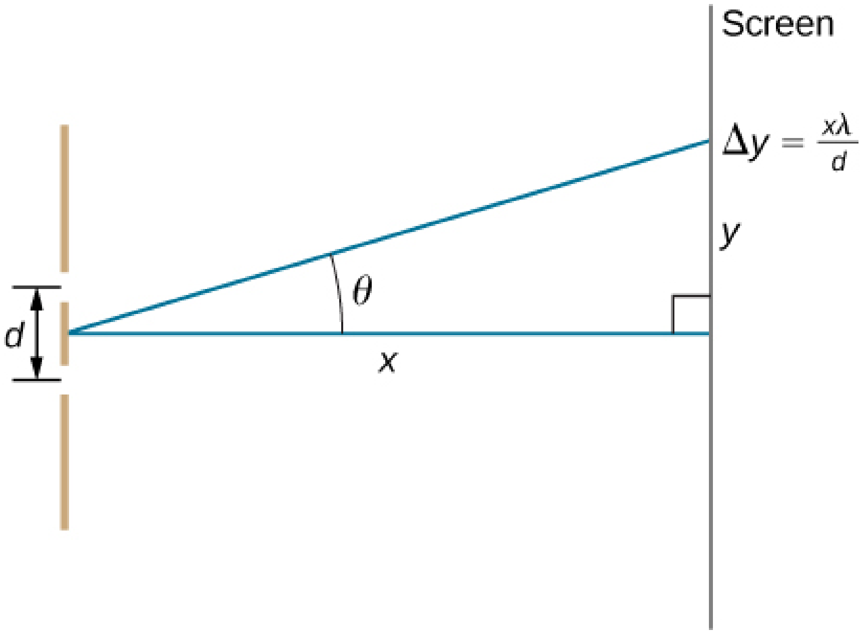
Solution:

a. 20.3°; b. 4.98°; c. 5.76, the highest order is $m = 5$.

Exercise:

Problem:

Shown below is a double slit located a distance *x* from a screen, with the distance from the center of the screen given by *y*. When the distance *d* between the slits is relatively large, numerous bright spots appear, called fringes. Show that, for small angles (where $\sin\theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = x\lambda/d$



Exercise:

Problem:

Using the result of the preceding problem, (a) calculate the distance between fringes for 633-nm light falling on double slits separated by 0.0800 mm, located 3.00 m from a screen. (b) What would be the distance between fringes if the entire apparatus were submersed in water, whose index of refraction is 1.33?

Solution:

a. 2.37 cm; b. 1.78 cm

Exercise:

Problem:

Using the result of the problem two problems prior, find the wavelength of light that produces fringes 7.50 mm apart on a screen 2.00 m from double slits separated by 0.120 mm.

Exercise:

Problem:

In a double-slit experiment, the fifth maximum is 2.8 cm from the central maximum on a screen that is 1.5 m away from the slits. If the slits are 0.15 mm apart, what is the wavelength of the light being used?

Solution:

560 nm

Exercise:

Problem:

The source in Young’s experiment emits at two wavelengths. On the viewing screen, the fourth maximum for one wavelength is located at the same spot as the fifth maximum for the other wavelength. What is the ratio of the two wavelengths?

Exercise:

Problem:

If 500-nm and 650-nm light illuminates two slits that are separated by 0.50 mm, how far apart are the second-order maxima for these two wavelengths on a screen 2.0 m away?

Solution:

1.2 mm

Exercise:

Problem:

Red light of wavelength of 700 nm falls on a double slit separated by 400 nm. (a) At what angle is the first-order maximum in the diffraction pattern? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

fringes
bright and dark patterns of interference

order
integer *m* used in the equations for constructive and destructive interference for a double slit

Multiple-Slit Interference

By the end of this section, you will be able to:

- Describe the locations and intensities of secondary maxima for multiple-slit interference

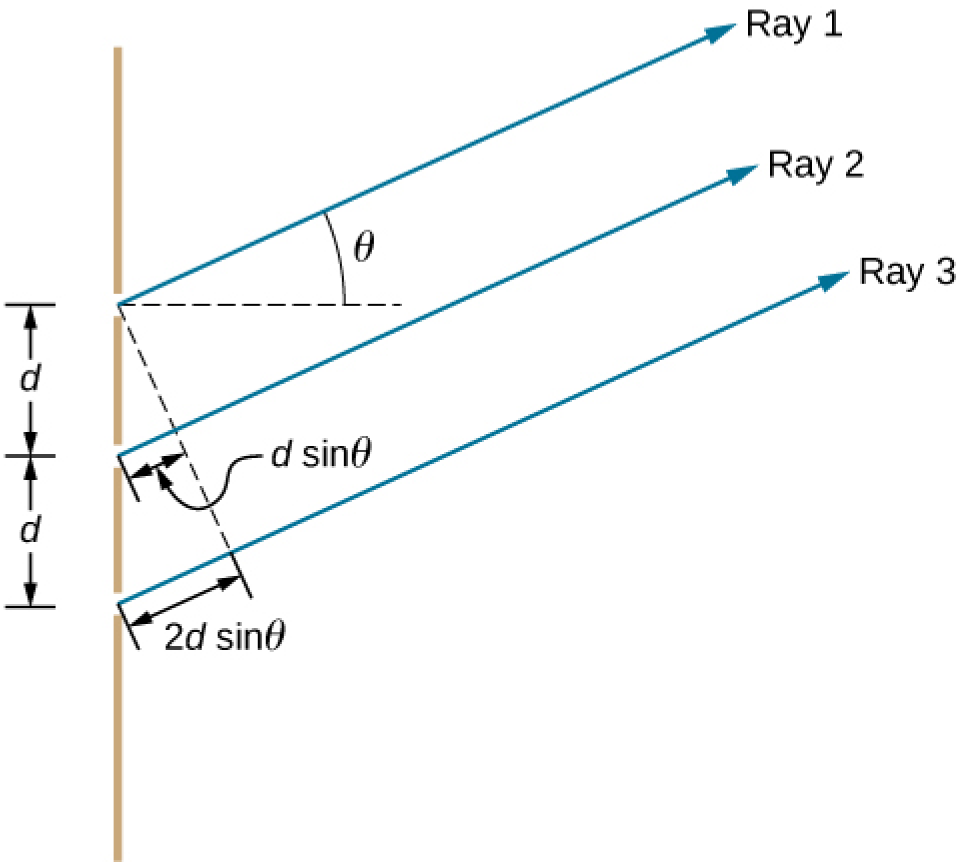
Analyzing the interference of light passing through two slits lays out the theoretical framework of interference and gives us a historical insight into Thomas Young’s experiments. However, much of the modern-day application of slit interference uses not just two slits but many, approaching infinity for practical purposes. The key optical element is called a diffraction grating, an important tool in optical analysis, which we discuss in detail in Diffraction. Here, we start the analysis of multiple-slit interference by taking the results from our analysis of the double slit ($N = 2$) and extending it to configurations with three, four, and much larger numbers of slits.

[link] shows the simplest case of multiple-slit interference, with three slits, or $N = 3$. The spacing between slits is d, and the path length difference between adjacent slits is $d \sin \theta$, same as the case for the double slit. What is new is that the path length difference for the first and the third slits is $2d \sin \theta$. The condition for constructive interference is the same as for the double slit, that is

Equation:

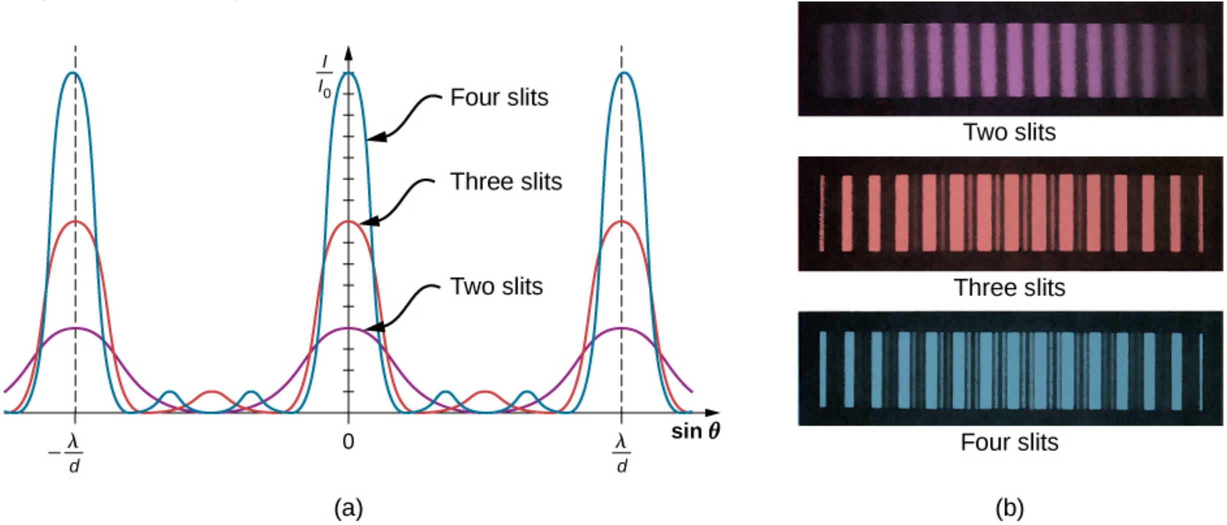
$$d \sin \theta = m\lambda.$$

When this condition is met, $2d \sin \theta$ is automatically a multiple of λ, so all three rays combine constructively, and the bright fringes that occur here are called **principal maxima**. But what happens when the path length difference between adjacent slits is only $\lambda/2$? We can think of the first and second rays as interfering destructively, but the third ray remains unaltered. Instead of obtaining a dark fringe, or a minimum, as we did for the double slit, we see a **secondary maximum** with intensity lower than the principal maxima.



Interference with three slits. Different pairs of emerging rays can combine constructively or destructively at the same time, leading to secondary maxima.

In general, for N slits, these secondary maxima occur whenever an unpaired ray is present that does not go away due to destructive interference. This occurs at $(N-2)$ evenly spaced positions between the principal maxima. The amplitude of the electromagnetic wave is correspondingly diminished to $1/N$ of the wave at the principal maxima, and the light intensity, being proportional to the square of the wave amplitude, is diminished to $1/N^2$ of the intensity compared to the principal maxima. As [link] shows, a dark fringe is located between every maximum (principal or secondary). As N grows larger and the number of bright and dark fringes increase, the widths of the maxima become narrower due to the closely located neighboring dark fringes. Because the total amount of light energy remains unaltered, narrower maxima require that each maximum reaches a correspondingly higher intensity.



(a)

(b)

Interference fringe patterns for two, three and four slits. As the number of slits increases, more secondary maxima appear, but the principal maxima become brighter and narrower. (a) Graph and (b) photographs of fringe patterns.

Summary

- Interference from multiple slits ($N > 2$) produces principal as well as secondary maxima.
- As the number of slits is increased, the intensity of the principal maxima increases and the width decreases.

Problems

Exercise:

Problem:

Ten narrow slits are equally spaced 0.25 mm apart and illuminated with yellow light of wavelength 580 nm. (a) What are the angular positions of the third and fourth principal maxima? (b) What is the separation of these maxima on a screen 2.0 m from the slits?

Solution:

a. $0.40^\circ, 0.53^\circ$; b. $4.6 \times 10^{-3}\,\text{m}$

Exercise:

Problem:

The width of bright fringes can be calculated as the separation between the two adjacent dark fringes on either side. Find the angular widths of the third- and fourth-order bright fringes from the preceding problem.

Exercise:

Problem:

For a three-slit interference pattern, find the ratio of the peak intensities of a secondary maximum to a principal maximum.

Solution:

1:9

Exercise:

Problem:

What is the angular width of the central fringe of the interference pattern of (a) 20 slits separated by $d = 2.0 \times 10^{-3}$mm? (b) 50 slits with the same separation? Assume that $\lambda = 600\,\text{nm}$.

Glossary

principal maximum
brightest interference fringes seen with multiple slits

secondary maximum
bright interference fringes of intensity lower than the principal maxima

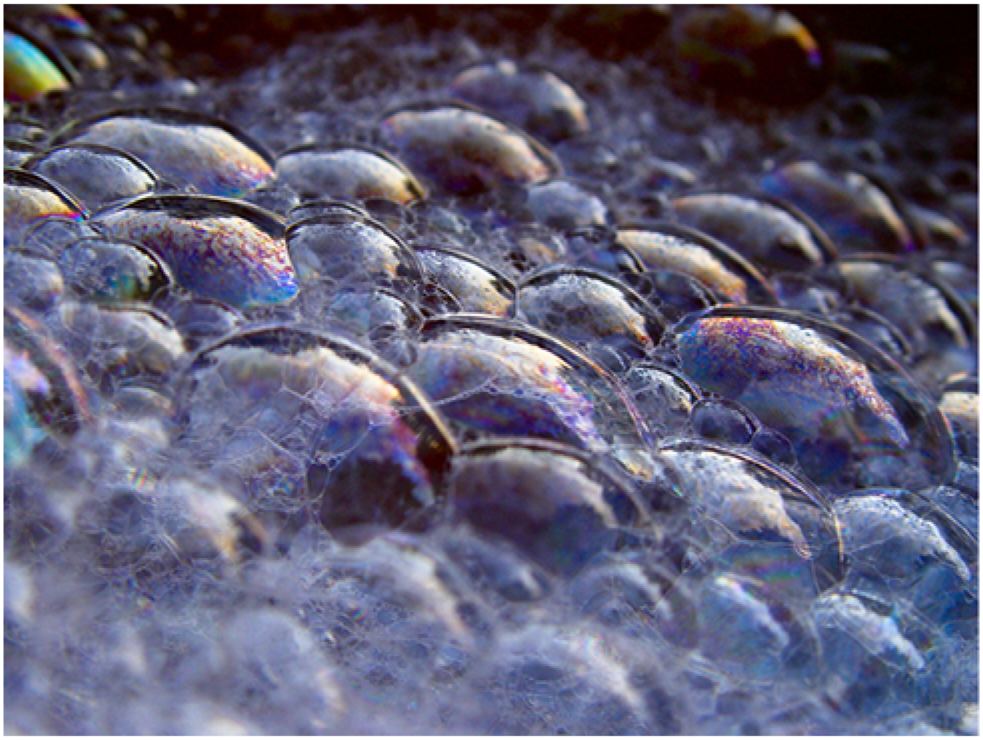
Interference in Thin Films

By the end of this section, you will be able to:

- Describe the phase changes that occur upon reflection
- Describe fringes established by reflected rays of a common source
- Explain the appearance of colors in thin films

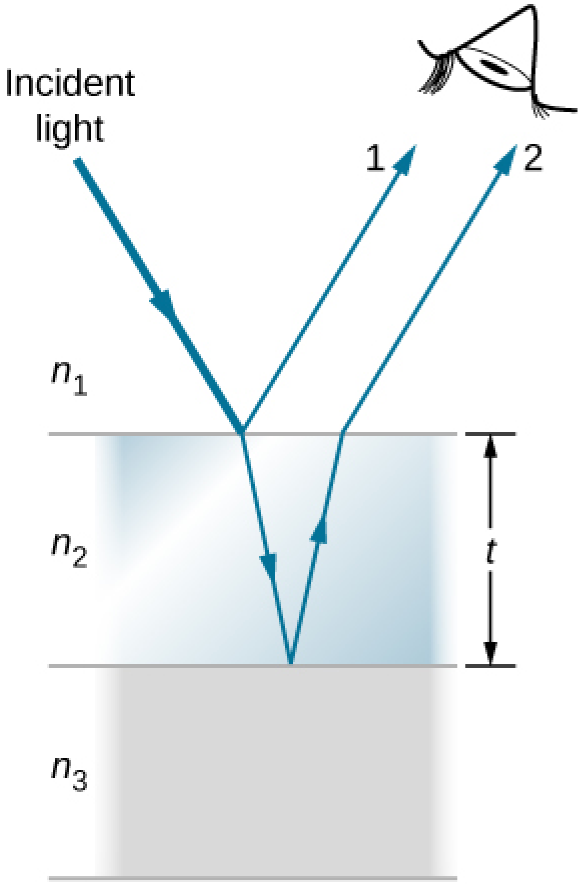
The bright colors seen in an oil slick floating on water or in a sunlit soap bubble are caused by interference. The brightest colors are those that interfere constructively. This interference is between light reflected from different surfaces of a thin film; thus, the effect is known as **thin-film interference**.

As we noted before, interference effects are most prominent when light interacts with something having a size similar to its wavelength. A thin film is one having a thickness *t* smaller than a few times the wavelength of light, λ. Since color is associated indirectly with λ and because all interference depends in some way on the ratio of λ to the size of the object involved, we should expect to see different colors for different thicknesses of a film, as in [link].



These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson)

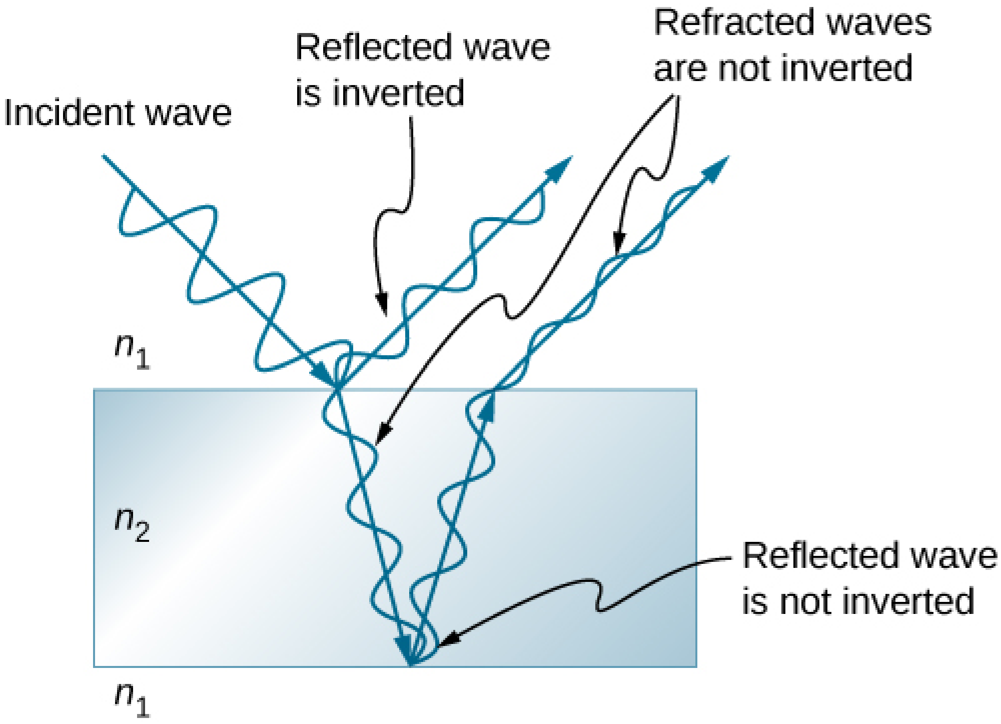
What causes thin-film interference? [link] shows how light reflected from the top and bottom surfaces of a film can interfere. Incident light is only partially reflected from the top surface of the film (ray 1). The remainder enters the film and is itself partially reflected from the bottom surface. Part of the light reflected from the bottom surface can emerge from the top of the film (ray 2) and interfere with light reflected from the top (ray 1). The ray that enters the film travels a greater distance, so it may be in or out of phase with the ray reflected from the top. However, consider for a moment, again, the bubbles in [link]. The bubbles are darkest where they are thinnest. Furthermore, if you observe a soap bubble carefully, you will note it gets dark at the point where it breaks. For very thin films, the difference in path lengths of rays 1 and 2 in [link] is negligible, so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection, as discussed next.



Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface and emerges as ray 2. These rays interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

Changes in Phase due to Reflection

We saw earlier (Waves) that reflection of mechanical waves can involve a 180° phase change. For example, a traveling wave on a string is inverted (i.e., a 180° phase change) upon reflection at a boundary to which a heavier string is tied. However, if the second string is lighter (or more precisely, of a lower linear density), no inversion occurs. Light waves produce the same effect, but the deciding parameter for light is the index of refraction. Light waves undergo a 180° or π radians phase change upon reflection at an interface beyond which is a medium of higher index of refraction. No phase change takes place when reflecting from a medium of lower refractive index ([link]). Because of the periodic nature of waves, this phase change or inversion is equivalent to $\pm\lambda/2$ in distance travelled, or path length. Both the path length and refractive indices are important factors in thin-film interference.


Incident wave
Reflected wave is inverted
Refracted waves are not inverted
n_1
n_2
Reflected wave is not inverted
n_1

Reflection at an interface for light traveling from a medium with index of refraction n_1 to a medium with index of refraction n_2, $n_1 < n_2$, causes the phase of the wave to change by π radians.

If the film in [link] is a **soap bubble** (essentially water with air on both sides), then a phase shift of $\lambda/2$ occurs for ray 1 but not for ray 2. Thus, when the film is very thin and the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference occurs at all wavelengths. Thus, the soap bubble is dark here. The thickness of the film relative to the wavelength of light is the other crucial factor in thin-film interference. Ray 2 in [link] travels a greater distance than ray 1. For light incident perpendicular to the surface, ray 2 travels a distance approximately 2*t* farther than ray 1. When this distance is an integral or half-integral multiple of the wavelength in the medium $(\lambda_n = \lambda/n$, where λ is the wavelength in vacuum and *n* is the index of refraction), constructive or destructive interference occurs, depending also on whether there is a phase change in either ray.

Example:

Calculating the Thickness of a Nonreflective Lens Coating

Sophisticated cameras use a series of several lenses. Light can reflect from the surfaces of these various lenses and degrade image clarity. To limit these reflections, lenses are coated with a thin layer of magnesium fluoride, which causes destructive thin-film interference. What is the thinnest this film can be, if its index of refraction is 1.38 and it is designed to limit the reflection of 550-nm light, normally the most intense visible wavelength? Assume the index of refraction of the glass is 1.52.

Strategy

Refer to [link] and use $n_1 = 1.00$ for air, $n_2 = 1.38$, and $n_3 = 1.52$. Both ray 1 and ray 2 have a $\lambda/2$ shift upon reflection. Thus, to obtain destructive interference, ray 2 needs to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is 2*t*.

Solution

To obtain destructive interference here,

Equation:

$$2t = \frac{\lambda_{n2}}{2}$$

where λ_{n2} is the wavelength in the film and is given by $\lambda_{n2} = \lambda / n_2$. Thus,

Equation:

$$2t = \frac{\lambda / n_2}{2}.$$

Solving for *t* and entering known values yields

Equation:

$$t = \frac{\lambda / n_2}{4} = \frac{(500\,\text{nm})/1.38}{4} = 90.6\,\text{nm}.$$

Significance

Films such as the one in this example are most effective in producing destructive interference when the thinnest layer is used, since light over a broader range of incident angles is reduced in intensity. These films are called nonreflective coatings; this is only an approximately correct description, though, since other wavelengths are only partially cancelled. Nonreflective coatings are also used in car windows and sunglasses.

Combining Path Length Difference with Phase Change

Thin-film interference is most constructive or most destructive when the path length difference for the two rays is an integral or half-integral wavelength. That is, for rays incident perpendicularly,

Equation:

$$2t = \lambda_n, 2\lambda_n, 3\lambda_n, \ldots \text{ or } 2t = \lambda_n/2, 3\lambda_n/2, 5\lambda_n/2, \ldots.$$

To know whether interference is constructive or destructive, you must also determine if there is a phase change upon reflection. Thin-film interference thus depends on film thickness, the wavelength of light, and the refractive indices. For white light incident on a film that varies in thickness, you can observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example:

Soap Bubbles

(a) What are the three smallest thicknesses of a soap bubble that produce constructive interference for red light with a wavelength of 650 nm? The index of refraction of soap is taken to be the same as that of water. (b) What three smallest thicknesses give destructive interference?

Strategy

Use [link] to visualize the bubble, which acts as a thin film between two layers of air. Thus $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a $\lambda/2$ shift for ray 1 reflected from the top surface of the bubble and no shift for ray 2 reflected from the bottom surface. To get constructive interference, then, the path length difference (2*t*) must be a half-integral multiple of the wavelength—the first three being $\lambda_n/2, 3\lambda_n/2$, and $5\lambda_n/2$. To get destructive interference, the path length difference must be an integral multiple of the wavelength—the first three being 0, λ_n, and $2\lambda_n$.

Solution

a. Constructive interference occurs here when

Equation:

$$2t_{\text{c}} = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \frac{5\lambda_n}{2}, \ldots .$$

Thus, the smallest constructive thickness t_{c} is

Equation:

$$t_{\text{c}} = \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{(650\,\text{nm})/1.333}{4} = 122\,\text{nm}.$$

The next thickness that gives constructive interference is $t'_{\text{c}} = 3\lambda_n/4$, so that

Equation:

$$t'_{\mathrm{c}} = 366\,\mathrm{nm}.$$

Finally, the third thickness producing constructive interference is $t'_{\mathrm{c}} = 5\lambda_n/4$, so that

Equation:

$$t'_{\mathrm{c}} = 610\,\mathrm{nm}.$$

b. For destructive interference, the path length difference here is an integral multiple of the wavelength. The first occurs for zero thickness, since there is a phase change at the top surface, that is,

Equation:

$$t_{\mathrm{d}} = 0,$$

the very thin (or negligibly thin) case discussed above. The first non-zero thickness producing destructive interference is

Equation:

$$2t'_{\mathrm{d}} = \lambda_n.$$

Substituting known values gives

Equation:

$$t'_{\mathrm{d}} = \frac{\lambda}{2} = \frac{\lambda/n}{2} = \frac{(650\,\mathrm{nm})/1.333}{2} = 244\,\mathrm{nm}.$$

Finally, the third destructive thickness is $2t''_{\mathrm{d}} = 2\lambda_n$, so that

Equation:

$$t''_{\mathrm{d}} = \lambda_n = \frac{\lambda}{n} = \frac{650\,\mathrm{nm}}{1.333} = 488\,\mathrm{nm}.$$

Significance

If the bubble were illuminated with pure red light, we would see bright and dark bands at very uniform increases in thickness. First would be a dark band at 0 thickness, then bright at 122 nm thickness, then dark at 244 nm, bright at 366 nm, dark at 488 nm, and bright at 610 nm. If the bubble varied smoothly in thickness, like a smooth wedge, then the bands would be evenly spaced.

Note:
Exercise:

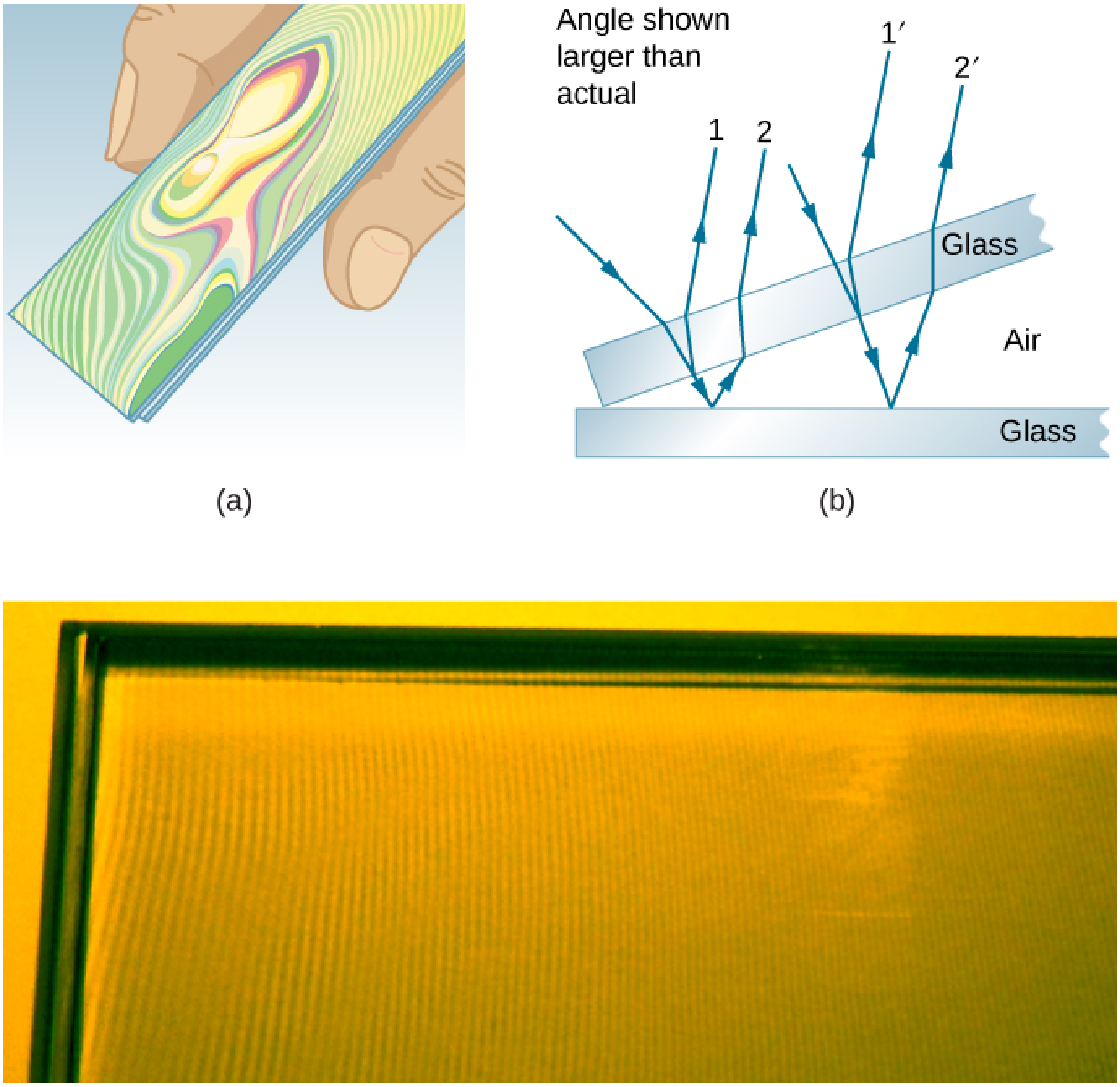
Problem:

Check Your Understanding Going further with [link], what are the next two thicknesses of soap bubble that would lead to (a) constructive interference, and (b) destructive interference?

Solution:

a. 853 nm, 1097 nm; b. 731 nm, 975 nm

Another example of thin-film interference can be seen when microscope slides are separated (see [link]). The slides are very flat, so that the wedge of air between them increases in thickness very uniformly. A phase change occurs at the second surface but not the first, so a dark band forms where the slides touch. The rainbow colors of constructive interference repeat, going from violet to red again and again as the distance between the slides increases. As the layer of air increases, the bands become more difficult to see, because slight changes in incident angle have greater effects on path length differences. If monochromatic light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.


Angle shown larger than actual
1
2
1′
2′
Glass
Air
Glass

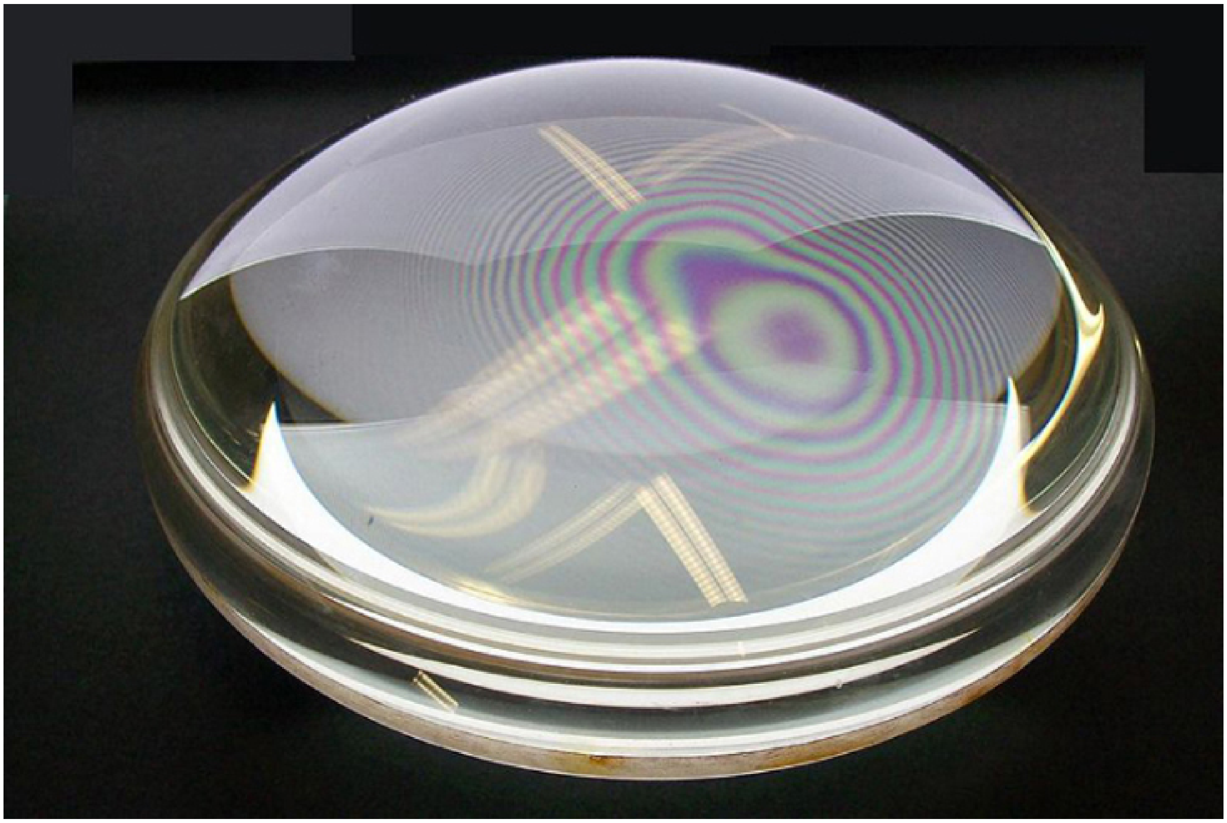
(a)

(b)

(c)

(a) The rainbow-color bands are produced by thin-film interference in the air between the two glass slides. (b) Schematic of the paths taken by rays in the wedge of air between the slides. (c) If the air wedge is illuminated with monochromatic light, bright and dark bands are obtained rather than repeating rainbow colors.

An important application of thin-film interference is found in the manufacturing of optical instruments. A lens or mirror can be compared with a master as it is being ground, allowing it to be shaped to an accuracy of less than a wavelength over its entire surface. [link] illustrates the phenomenon called **Newton’s rings**, which occurs when the plane surfaces of two lenses are placed together. (The circular bands are called Newton’s rings because Isaac Newton described them and their use in detail. Newton did not discover them; Robert Hooke did, and Newton did not believe they were due to the wave character of light.) Each successive ring of a given color indicates an increase of only half a wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, no rings appear.



“Newton’s rings” interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert)

Thin-film interference has many other applications, both in nature and in manufacturing. The wings of certain moths and butterflies have nearly iridescent colors due to thin-film interference. In addition to pigmentation, the wing's color is affected greatly by constructive interference of certain wavelengths reflected from its film-coated surface. Some car manufacturers offer special paint jobs that use thin-film interference to produce colors that change with angle. This expensive option is based on variation of thin-film path length differences with angle. Security features on credit cards, banknotes, driving licenses, and similar items prone to forgery use thin-film interference, diffraction gratings, or holograms. As early as 1998, Australia led the way with dollar bills printed on polymer with a diffraction grating security feature, making the currency difficult to forge. Other countries, such as Canada, New Zealand, and Taiwan, are using similar technologies, while US currency includes a thin-film interference effect.

Summary

- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a $\lambda/2$ shift) occurs.
- Thin-film interference occurs between the light reflected from the top and bottom surfaces of a film. In addition to the path length difference, there can be a phase change.

Conceptual Questions

Exercise:

Problem:

What effect does increasing the wedge angle have on the spacing of interference fringes? If the wedge angle is too large, fringes are not observed. Why?

Exercise:

Problem:

How is the difference in paths taken by two originally in-phase light waves related to whether they interfere constructively or destructively? How can this be affected by reflection? By refraction?

Solution:

Differing path lengths result in different phases at destination resulting in constructive or destructive interference accordingly. Reflection can cause a 180° phase change, which also affects how waves interfere. Refraction into another medium changes the wavelength inside that medium such that a wave can emerge from the medium with a different phase compared to another wave that travelled the same distance in a different medium.

Exercise:

Problem:

Is there a phase change in the light reflected from either surface of a contact lens floating on a person’s tear layer? The index of refraction of the lens is about 1.5, and its top surface is dry.

Exercise:

Problem:

In placing a sample on a microscope slide, a glass cover is placed over a water drop on the glass slide. Light incident from above can reflect from the top and bottom of the glass cover and from the glass slide below the water drop. At which surfaces will there be a phase change in the reflected light?

Solution:

Phase changes occur upon reflection at the top of glass cover and the top of glass slide only.

Exercise:

Problem:

Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.

Exercise:

Problem:

While contemplating the food value of a slice of ham, you notice a rainbow of color reflected from its moist surface. Explain its origin.

Solution:

The surface of the ham being moist means there is a thin layer of fluid, resulting in thin-film interference. Because the exact thickness of the film varies across the piece of ham, which is illuminated by white light, different wavelengths produce bright fringes at different locations, resulting in rainbow colors.

Exercise:

Problem:

An inventor notices that a soap bubble is dark at its thinnest and realizes that destructive interference is taking place for all wavelengths. How could she use this knowledge to make a nonreflective coating for lenses that is effective at all wavelengths? That is, what limits would there be on the index of refraction and thickness of the coating? How might this be impractical?

Exercise:

Problem:

A nonreflective coating like the one described in [link] works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.

Solution:

Other wavelengths will not generally satisfy $t = \frac{\lambda/n}{4}$ for the same value of *t* so reflections will result in completely destructive interference. For an incidence angle θ, the path length inside the coating will be increased by a factor $1/\cos\theta$ so the new condition for destructive interference becomes $\frac{t}{\cos\theta} = \frac{\lambda/n}{4}$.

Exercise:

Problem:

Why is it much more difficult to see interference fringes for light reflected from a thick piece of glass than from a thin film? Would it be easier if monochromatic light were used?

Problems

Exercise:

Problem:

A soap bubble is 100 nm thick and illuminated by white light incident perpendicular to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

Solution:

532 nm (green)

Exercise:

Problem:

An oil slick on water is 120 nm thick and illuminated by white light incident perpendicular to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

Exercise:

Problem:

Calculate the minimum thickness of an oil slick on water that appears blue when illuminated by white light perpendicular to its surface. Take the blue wavelength to be 470 nm and the index of refraction of oil to be 1.40.

Solution:

$8.39 \times 10^{-8}\,\text{m} = 83.9\,\text{nm}$

Exercise:

Problem:

Find the minimum thickness of a soap bubble that appears red when illuminated by white light perpendicular to its surface. Take the wavelength to be 680 nm, and assume the same index of refraction as water.

Exercise:

Problem:

A film of soapy water ($n = 1.33$) on top of a plastic cutting board has a thickness of 233 nm. What color is most strongly reflected if it is illuminated perpendicular to its surface?

Solution:

620 nm (orange)

Exercise:

Problem:

What are the three smallest non-zero thicknesses of soapy water ($n = 1.33$) on Plexiglas if it appears green (constructively reflecting 520-nm light) when illuminated perpendicularly by white light?

Exercise:

Problem:

Suppose you have a lens system that is to be used primarily for 700-nm red light. What is the second thinnest coating of fluorite (magnesium fluoride) that would be nonreflective for this wavelength?

Solution:

380 nm

Exercise:

Problem:

(a) As a soap bubble thins it becomes dark, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the bubble can be and appear dark at all visible wavelengths? Assume the same index of refraction as water. (b) Discuss the fragility of the film considering the thickness found.

Exercise:

Problem:

To save money on making military aircraft invisible to radar, an inventor decides to coat them with a nonreflective material having an index of refraction of 1.20, which is between that of air and the surface of the plane. This, he reasons, should be much cheaper than designing Stealth bombers. (a) What thickness should the coating be to inhibit the reflection of 4.00-cm wavelength radar? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. Assuming *n* for the plane is greater than 1.20, then there are two phase changes: 0.833 cm. b. It is too thick, and the plane would be too heavy. c. It is unreasonable to think the layer of material could be any thickness when used on a real aircraft.

Glossary

Newton’s rings
: circular interference pattern created by interference between the light reflected off two surfaces as a result of a slight gap between them

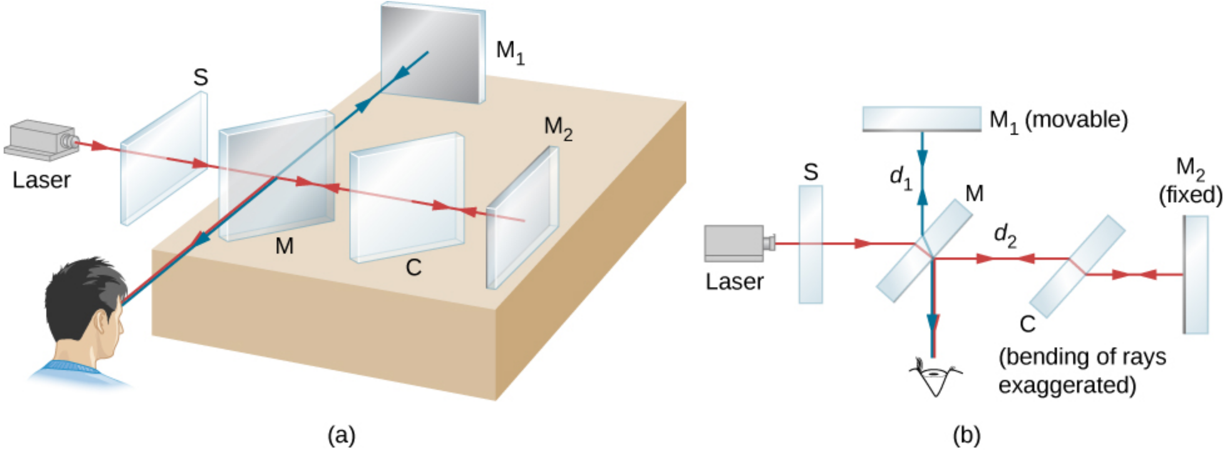
thin-film interference
: interference between light reflected from different surfaces of a thin film

The Michelson Interferometer

By the end of this section, you will be able to:

- Explain changes in fringes observed with a Michelson interferometer caused by mirror movements
- Explain changes in fringes observed with a Michelson interferometer caused by changes in medium

The Michelson **interferometer** (invented by the American physicist Albert A. Michelson, 1852–1931) is a precision instrument that produces interference fringes by splitting a light beam into two parts and then recombining them after they have traveled different optical paths. [link] depicts the interferometer and the path of a light beam from a single point on the extended source S, which is a ground-glass plate that diffuses the light from a monochromatic lamp of wavelength λ_0. The beam strikes the half-silvered mirror M, where half of it is reflected to the side and half passes through the mirror. The reflected light travels to the movable plane mirror M_1, where it is reflected back through M to the observer. The transmitted half of the original beam is reflected back by the stationary mirror M_2 and then toward the observer by M.



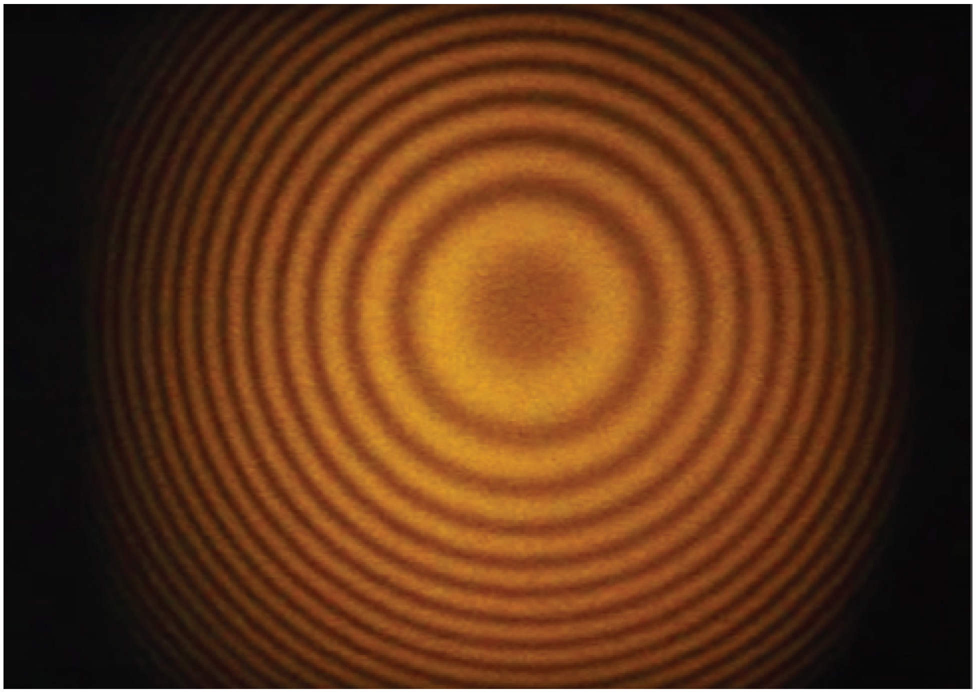
(a) The Michelson interferometer. The extended light source is a ground-glass plate that diffuses the light from a laser. (b) A planar view of the interferometer.

Because both beams originate from the same point on the source, they are coherent and therefore interfere. Notice from the figure that one beam passes through M three times and the other only once. To ensure that both beams traverse the same thickness of glass, a compensator plate C of transparent glass is placed in the arm containing M_2. This plate is a duplicate of M (without the silvering) and is usually cut from the same piece of glass used to produce M. With the compensator in place, any phase difference between the two beams is due solely to the difference in the distances they travel.

The path difference of the two beams when they recombine is $2d_1 - 2d_2$, where d_1 is the distance between M and M_1, and d_2 is the distance between M and M_2. Suppose this path difference is an integer number of wavelengths $m\lambda_0$. Then, constructive interference occurs and a bright image of the point on the source is seen at the observer. Now the light from any other point on the source whose two beams have this same path difference also undergoes constructive interference and produces a bright image. The collection of these point images is a bright fringe corresponding to a path difference of $m\lambda_0$ ([link]). When M_1 is moved a distance $\Delta d = \lambda_0/2$, this path difference changes by λ_0, and each fringe moves to the position previously occupied by an adjacent fringe. Consequently, by counting the number of fringes *m* passing a given point as M_1 is moved, an observer can measure minute displacements that are accurate to a fraction of a wavelength, as shown by the relation

Equation:

$$\Delta d = m\frac{\lambda_0}{2}.$$



Fringes produced with a Michelson interferometer.
(credit: “SILLAGESvideos”/YouTube)

Example:

Precise Distance Measurements by Michelson Interferometer

A red laser light of wavelength 630 nm is used in a Michelson interferometer. While keeping the mirror M_1 fixed, mirror M_2 is moved. The fringes are found to move past a fixed cross-hair in the viewer. Find the distance the mirror M_2 is moved for a single fringe to move past the reference line.

Strategy

Refer to [link] for the geometry. We use the result of the Michelson interferometer interference condition to find the distance moved, Δd.

Solution

For a 630-nm red laser light, and for each fringe crossing $(m = 1)$, the distance traveled by M_2 if you keep M_1 fixed is

Equation:

$$\Delta d = m\frac{\lambda_0}{2} = 1 \times \frac{630\ \text{nm}}{2} = 315\ \text{nm} = 0.315\ \mu\text{m}.$$

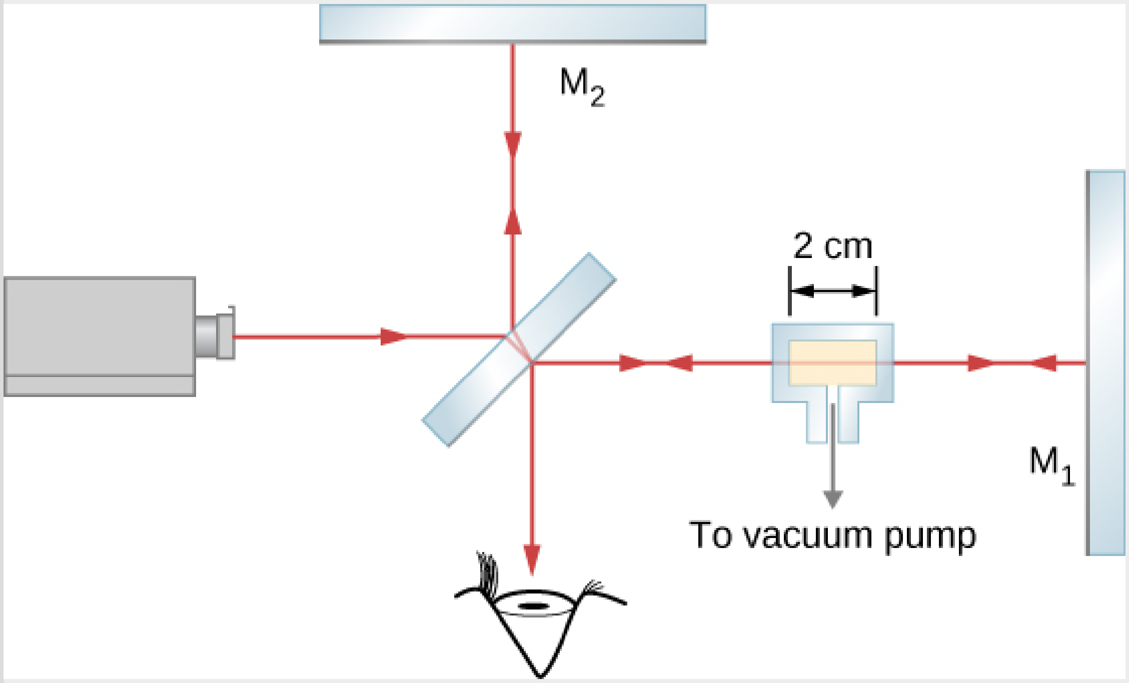
Significance

An important application of this measurement is the definition of the standard meter. As mentioned in Units and Measurement, the length of the standard meter was once defined as the mirror displacement in a Michelson interferometer corresponding to 1,650,763.73 wavelengths of the particular fringe of krypton-86 in a gas discharge tube.

Example:

Measuring the Refractive Index of a Gas

In one arm of a Michelson interferometer, a glass chamber is placed with attachments for evacuating the inside and putting gases in it. The space inside the container is 2 cm wide. Initially, the container is empty. As gas is slowly let into the chamber, you observe that dark fringes move past a reference line in the field of observation. By the time the chamber is filled to the desired pressure, you have counted 122 fringes move past the reference line. The wavelength of the light used is 632.8 nm. What is the refractive index of this gas?



Strategy

The $m = 122$ fringes observed compose the difference between the number of wavelengths that fit within the empty chamber (vacuum) and the number of wavelengths that fit within the same chamber when it is gas-filled. The wavelength in the filled chamber is shorter by a factor of n, the index of refraction.

Solution

The ray travels a distance $t = 2\ \text{cm}$ to the right through the glass chamber and another distance t to the left upon reflection. The total travel is $L = 2t$. When empty, the number of wavelengths that fit in this chamber is

Equation:

$$N_0 = \frac{L}{\lambda_0} = \frac{2t}{\lambda_0}$$

where $\lambda_0 = 632.8\ \text{nm}$ is the wavelength in vacuum of the light used. In any other medium, the wavelength is $\lambda = \lambda_0/n$ and the number of wavelengths that fit in the gas-filled chamber is

Equation:

$$N = \frac{L}{\lambda} = \frac{2t}{\lambda_0/n}.$$

The number of fringes observed in the transition is

Equation:

$$\begin{aligned} m &= N - N_0, \\ &= \tfrac{2t}{\lambda_0/n} - \tfrac{2t}{\lambda_0}, \\ &= \tfrac{2t}{\lambda_0}(n-1). \end{aligned}$$

Solving for $(n - 1)$ gives

Equation:

$$n - 1 = m\left(\frac{\lambda_0}{2t}\right) = 122\left(\frac{632.8\ \times\ 10^{-9}\,\text{m}}{2(2\ \times\ 10^{-2}\,\text{m})}\right) = 0.0019$$

and $n = 1.0019$.

Significance

The indices of refraction for gases are so close to that of vacuum, that we normally consider them equal to 1. The difference between 1 and 1.0019 is so small that measuring it requires a correspondingly sensitive technique such as interferometry. We cannot, for example, hope to measure this value using techniques based simply on Snell's law.

Note:

Exercise:

Problem:

Check Your Understanding Although *m*, the number of fringes observed, is an integer, which is often regarded as having zero uncertainty, in practical terms, it is all too easy to lose track when counting fringes. In [link], if you estimate that you might have missed as many as five fringes when you reported $m = 122$ fringes, (a) is the value for the index of refraction worked out in [link] too large or too small? (b) By how much?

Solution:

a. too small; b. up to 8×10^{-5}

Note:

Problem-Solving Strategy: Wave Optics

Step 1. *Examine the situation to determine that interference is involved*. Identify whether slits, thin films, or interferometers are considered in the problem.

Step 2. *If slits are involved*, note that diffraction gratings and double slits produce very similar interference patterns, but that gratings have narrower (sharper) maxima. Single-slit patterns are characterized by a large central maximum and smaller maxima to the sides.

Step 3. *If thin-film interference or an interferometer is involved, take note of the path length difference between the two rays that interfere*. Be certain to use the wavelength in the medium involved, since it differs from the wavelength in vacuum. Note also that there is an additional $\lambda/2$ phase shift when light reflects from a medium with a greater index of refraction.
Step 4. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. A written list is useful. Draw a diagram of the situation. Labeling the diagram is useful.
Step 5. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns)*.
Step 6. *Solve the appropriate equation for the quantity to be determined (the unknown) and enter the knowns*. Slits, gratings, and the Rayleigh limit involve equations.
Step 7. *For thin-film interference, you have constructive interference for a total shift that is an integral number of wavelengths. You have destructive interference for a total shift of a half-integral number of wavelengths*. Always keep in mind that crest to crest is constructive whereas crest to trough is destructive.
Step 8. *Check to see if the answer is reasonable: Does it make sense?* Angles in interference patterns cannot be greater than 90°, for example.

Summary

- When the mirror in one arm of the interferometer moves a distance of $\lambda/2$ each fringe in the interference pattern moves to the position previously occupied by the adjacent fringe.

Key Equations

Constructive	$\Delta l = m\lambda,$ for *m* = 0, ±1, ±2, ±3…

interference	
Destructive interference	$\Delta l = (m + \frac{1}{2})\lambda,$ for m = 0, ±1, ±2, ±3…
Path length difference for waves from two slits to a common point on a screen	$\Delta l = d \sin \theta$
Constructive interference	$d \sin \theta = m\lambda, \quad \text{for } m = 0, \pm 1, \pm 2, \pm 3, \ldots$
Destructive interference	$d \sin \theta = (m + \frac{1}{2})\lambda, \ \text{for } m \ = \ 0, \pm 1, \pm 2, \pm 3, \ldots$
Distance from central maximum to the *m*th bright fringe	$y_m = \frac{m\lambda D}{d}$
Displacement measured by a Michelson interferometer	$\Delta d = m \frac{\lambda_0}{2}$

Conceptual Questions

Exercise:

Problem:

Describe how a Michelson interferometer can be used to measure the index of refraction of a gas (including air).

Solution:

In one arm, place a transparent chamber to be filled with the gas. See [link].

Problems

Exercise:

Problem:

A Michelson interferometer has two equal arms. A mercury light of wavelength 546 nm is used for the interferometer and stable fringes are found. One of the arms is moved by $1.5\mu\text{m}$. How many fringes will cross the observing field?

Exercise:

Problem:

What is the distance moved by the traveling mirror of a Michelson interferometer that corresponds to 1500 fringes passing by a point of the observation screen? Assume that the interferometer is illuminated with a 606 nm spectral line of krypton-86.

Solution:

$4.55\,\times\,10^{-4}\,\text{m}$

Exercise:

Problem:

When the traveling mirror of a Michelson interferometer is moved $2.40 \times 10^{-5}\,\mathrm{m}$, 90 fringes pass by a point on the observation screen. What is the wavelength of the light used?

Exercise:

Problem:

In a Michelson interferometer, light of wavelength 632.8 nm from a He-Ne laser is used. When one of the mirrors is moved by a distance *D*, 8 fringes move past the field of view. What is the value of the distance *D*?

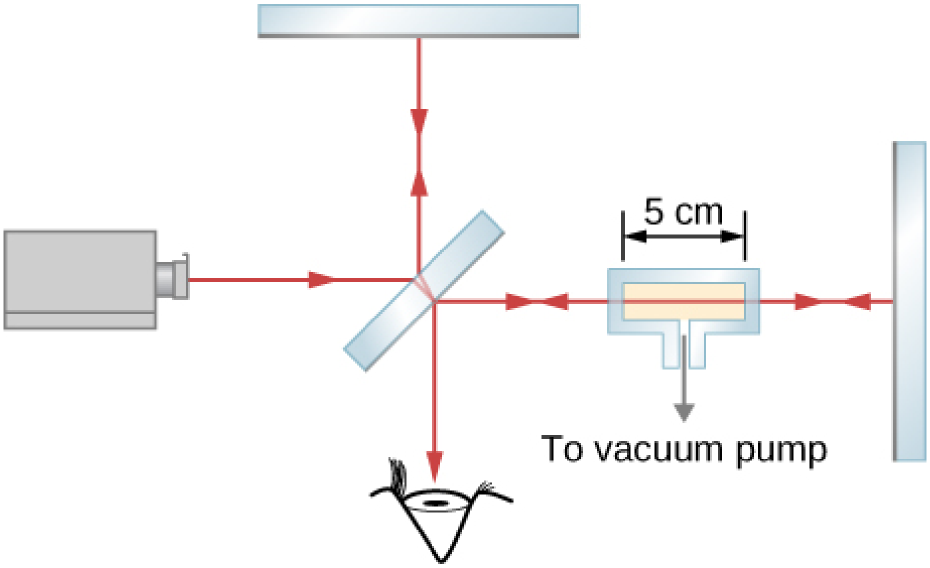
Solution:

$D = 2.53 \times 10^{-6}\,\mathrm{m}$

Exercise:

Problem:

A chamber 5.0 cm long with flat, parallel windows at the ends is placed in one arm of a Michelson interferometer (see below). The light used has a wavelength of 500 nm in a vacuum. While all the air is being pumped out of the chamber, 29 fringes pass by a point on the observation screen. What is the refractive index of the air?



Additional Problems

Exercise:

Problem:

For 600-nm wavelength light and a slit separation of 0.12 mm, what are the angular positions of the first and third maxima in the double slit interference pattern?

Solution:

0.29° and 0.86°

Exercise:

Problem:

If the light source in the preceding problem is changed, the angular position of the third maximum is found to be 0.57°. What is the wavelength of light being used now?

Exercise:

Problem:

Red light ($\lambda = 710.\ \text{nm}$) illuminates double slits separated by a distance $d = 0.150\ \text{mm}$. The screen and the slits are 3.00 m apart. (a) Find the distance on the screen between the central maximum and the third maximum. (b) What is the distance between the second and the fourth maxima?

Solution:

a. 4.26 cm; b. 2.84 cm

Exercise:

Problem:

Two sources as in phase and emit waves with $\lambda = 0.42\ \text{m}$. Determine whether constructive or destructive interference occurs at points whose distances from the two sources are (a) 0.84 and 0.42 m, (b) 0.21 and 0.42 m, (c) 1.26 and 0.42 m, (d) 1.87 and 1.45 m, (e) 0.63 and 0.84 m and (f) 1.47 and 1.26 m.

Exercise:

Problem:

Two slits $4.0\ \times\ 10^{-6}\ \text{m}$ apart are illuminated by light of wavelength 600 nm. What is the highest order fringe in the interference pattern?

Solution:

6

Exercise:

Problem:

Suppose that the highest order fringe that can be observed is the eighth in a double-slit experiment where 550-nm wavelength light is used. What is the minimum separation of the slits?

Exercise:

Problem:

The interference pattern of a He-Ne laser light $(\lambda = 632.9\,\text{nm})$ passing through two slits 0.031 mm apart is projected on a screen 10.0 m away. Determine the distance between the adjacent bright fringes.

Solution:

0.20 m

Exercise:

Problem:

Young's double-slit experiment is performed immersed in water ($n = 1.333$). The light source is a He-Ne laser, $\lambda = 632.9\,\text{nm}$ in vacuum. (a) What is the wavelength of this light in water? (b) What is the angle for the third order maximum for two slits separated by 0.100 mm.

Exercise:

Problem:

A double-slit experiment is to be set up so that the bright fringes appear 1.27 cm apart on a screen 2.13 m away from the two slits. The light source was wavelength 500 nm. What should be the separation between the two slits?

Solution:

0.0839 mm

Exercise:

Problem:

An effect analogous to two-slit interference can occur with sound waves, instead of light. In an open field, two speakers placed 1.30 m apart are powered by a single-function generator producing sine waves at 1200-Hz frequency. A student walks along a line 12.5 m away and parallel to the line between the speakers. She hears an alternating pattern of loud and quiet, due to constructive and destructive interference. What is (a) the wavelength of this sound and (b) the distance between the central maximum and the first maximum (loud) position along this line?

Exercise:

Problem:

A hydrogen gas discharge lamp emits visible light at four wavelengths, $\lambda =$ 410, 434, 486, and 656 nm. (a) If light from this lamp falls on a N slits separated by 0.025 mm, how far from the central maximum are the third maxima when viewed on a screen 2.0 m from the slits? (b) By what distance are the second and third maxima separated for $l = 486\,\text{nm}$?

Solution:

a. 9.8, 10.4, 11.7, and 15.7 cm; b. 3.9 cm

Exercise:

Problem:

Monochromatic light of frequency $5.5 \times 10^{14}\,\text{Hz}$ falls on 10 slits separated by 0.020 mm. What is the separation between the first and third maxima on a screen that is 2.0 m from the slits?

Exercise:

Problem:

Eight slits equally separated by 0.149 mm is uniformly illuminated by a monochromatic light at $\lambda = 523 \text{ nm}$. What is the width of the central principal maximum on a screen 2.35 m away?

Solution:

0.0575°

Exercise:

Problem:

Eight slits equally separated by 0.149 mm is uniformly illuminated by a monochromatic light at $\lambda = 523 \text{ nm}$. What is the intensity of a secondary maxima compared to that of the principal maxima?

Exercise:

Problem:

A transparent film of thickness 250 nm and index of refraction of 1.40 is surrounded by air. What wavelength in a beam of white light at near-normal incidence to the film undergoes destructive interference when reflected?

Solution:

700 nm

Exercise:

Problem:

An intensity minimum is found for 450 nm light transmitted through a transparent film $(n = 1.20)$ in air. (a) What is minimum thickness of the film? (b) If this wavelength is the longest for which the intensity minimum occurs, what are the next three lower values of λ for which this happens?

Exercise:

Problem:

A thin film with $n=1.32$ is surrounded by air. What is the minimum thickness of this film such that the reflection of normally incident light with $\lambda=500\text{ nm}$ is minimized?

Solution:

189 nm

Exercise:

Problem:

Repeat your calculation of the previous problem with the thin film placed on a flat glass ($n=1.50$) surface.

Exercise:

Problem:

After a minor oil spill, a think film of oil ($n=1.40$) of thickness 450 nm floats on the water surface in a bay. (a) What predominant color is seen by a bird flying overhead? (b) What predominant color is seen by a seal swimming underwater?

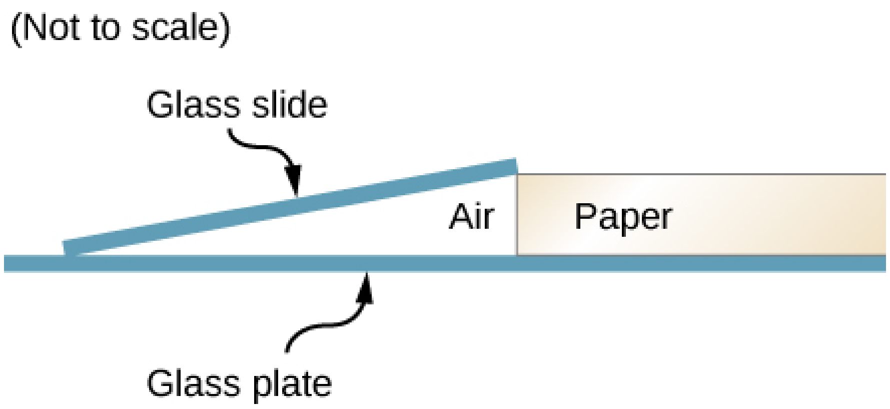
Solution:

a. green (504 nm); b. magenta (white minus green)

Exercise:

Problem:

A microscope slide 10 cm long is separated from a glass plate at one end by a sheet of paper. As shown below, the other end of the slide is in contact with the plate. The slide is illuminated from above by light from a sodium lamp ($\lambda=589\text{ nm}$), and 14 fringes per centimeter are seen along the slide. What is the thickness of the piece of paper?



Exercise:

Problem:

Suppose that the setup of the preceding problem is immersed in an unknown liquid. If 18 fringes per centimeter are now seen along the slide, what is the index of refraction of the liquid?

Solution:

1.29

Exercise:

Problem:

A thin wedge filled with air is produced when two flat glass plates are placed on top of one another and a slip of paper is inserted between them at one edge. Interference fringes are observed when monochromatic light falling vertically on the plates are seen in reflection. Is the first fringe near the edge where the plates are in contact a bright fringe or a dark fringe? Explain.

Exercise:

Problem:

Two identical pieces of rectangular plate glass are used to measure the thickness of a hair. The glass plates are in direct contact at one edge and a single hair is placed between them hear the opposite edge. When illuminated with a sodium lamp ($\lambda = 589\,\text{nm}$), the hair is seen between the 180th and 181st dark fringes. What are the lower and upper limits on the hair’s diameter?

Solution:

$52.7\ \mu\text{m}$ and $53.0\ \mu\text{m}$

Exercise:

Problem:

Two microscope slides made of glass are illuminated by monochromatic ($\lambda = 589\,\text{nm}$) light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a thin copper wire at the other end, forming a wedge of air. The diameter of the copper wire is $29.45\ \mu\text{m}$. How many bright fringes are seen across these slides?

Exercise:

Problem:

A good quality camera “lens” is actually a system of lenses, rather than a single lens, but a side effect is that a reflection from the surface of one lens can bounce around many times within the system, creating artifacts in the photograph. To counteract this problem, one of the lenses in such a system is coated with a thin layer of material ($n = 1.28$) on one side. The index of refraction of the lens glass is 1.68. What is the smallest thickness of the coating that reduces the reflection at 640 nm by destructive interference? (In other words, the coating’s effect is to be optimized for $\lambda = 640\,\text{nm}$.)

Solution:

160 nm

Exercise:

Problem:

Constructive interference is observed from directly above an oil slick for wavelengths (in air) 440 nm and 616 nm. The index of refraction of this oil is $n = 1.54$. What is the film's minimum possible thickness?

Exercise:

Problem:

A soap bubble is blown outdoors. What colors (indicate by wavelengths) of the reflected sunlight are seen enhanced? The soap bubble has index of refraction 1.36 and thickness 380 nm.

Solution:

413 nm and 689 nm

Exercise:

Problem:

A Michelson interferometer with a He-Ne laser light source ($\lambda = 632.8 \text{ nm}$) projects its interference pattern on a screen. If the movable mirror is caused to move by $8.54 \mu\text{m}$, how many fringes will be observed shifting through a reference point on a screen?

Exercise:

Problem:

An experimenter detects 251 fringes when the movable mirror in a Michelson interferometer is displaced. The light source used is a sodium lamp, wavelength 589 nm. By what distance did the movable mirror move?

Solution:

$73.9 \mu\text{m}$

Exercise:

Problem:

A Michelson interferometer is used to measure the wavelength of light put through it. When the movable mirror is moved by exactly 0.100 mm, the number of fringes observed moving through is 316. What is the wavelength of the light?

Exercise:

Problem:

A 5.08-cm-long rectangular glass chamber is inserted into one arm of a Michelson interferometer using a 633-nm light source. This chamber is initially filled with air $(n = 1.000293)$ at standard atmospheric pressure but the air is gradually pumped out using a vacuum pump until a near perfect vacuum is achieved. How many fringes are observed moving by during the transition?

Solution:

47

Exercise:

Problem:

Into one arm of a Michelson interferometer, a plastic sheet of thickness $75\ \mu\text{m}$ is inserted, which causes a shift in the interference pattern by 86 fringes. The light source has wavelength of 610 nm in air. What is the index of refraction of this plastic?

Exercise:

Problem:

The thickness of an aluminum foil is measured using a Michelson interferometer that has its movable mirror mounted on a micrometer. There is a difference of 27 fringes in the observed interference pattern when the micrometer clamps down on the foil compared to when the micrometer is empty. Calculate the thickness of the foil?

Solution:

$8.5\ \mu\text{m}$

Exercise:

Problem:

The movable mirror of a Michelson interferometer is attached to one end of a thin metal rod of length 23.3 mm. The other end of the rod is anchored so it does not move. As the temperature of the rod changes from $15\ ^\circ\text{C}$ to $25\ \text{C}$, a change of 14 fringes is observed. The light source is a He Ne laser, $\lambda = 632.8\ \text{nm}$. What is the change in length of the metal bar, and what is its thermal expansion coefficient?

Exercise:

Problem:

In a thermally stabilized lab, a Michelson interferometer is used to monitor the temperature to ensure it stays constant. The movable mirror is mounted on the end of a 1.00-m-long aluminum rod, held fixed at the other end. The light source is a He Ne laser, $\lambda = 632.8\ \text{nm}$. The resolution of this apparatus corresponds to the temperature difference when a change of just one fringe is observed. What is this temperature difference?

Solution:

0.013°C

Exercise:

Problem:

A 65-fringe shift results in a Michelson interferometer when a $42.0\text{-}\mu\text{m}$ film made of an unknown material is placed in one arm. The light source has wavelength 632.9 nm. Identify the material using the indices of refraction found in [link].

Challenge Problems

Exercise:

Problem:

Determine what happens to the double-slit interference pattern if one of the slits is covered with a thin, transparent film whose thickness is $\lambda/[2(n-1)]$, where λ is the wavelength of the incident light and n is the index of refraction of the film.

Solution:

Bright and dark fringes switch places.

Exercise:

Problem:

Fifty-one narrow slits are equally spaced and separated by 0.10 mm. The slits are illuminated by blue light of wavelength 400 nm. What is angular position of the twenty-fifth secondary maximum? What is its peak intensity in comparison with that of the primary maximum?

Exercise:

Problem:

A film of oil on water will appear dark when it is very thin, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the oil can be and appear dark at all visible wavelengths? Oil has an index of refraction of 1.40.

Solution:

The path length must be less than one-fourth of the shortest visible wavelength in oil. The thickness of the oil is half the path length, so it must be less than one-eighth of the shortest visible wavelength in oil. If we take 380 nm to be the shortest visible wavelength in air, 33.9 nm.

Exercise:

Problem:

[link] shows two glass slides illuminated by monochromatic light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a 0.100-mm-diameter hair at the other end, forming a wedge of air. (a) How far apart are the dark bands, if the slides are 7.50 cm long and 589-nm light is used? (b) Is there any difference if the slides are made from crown or flint glass? Explain.

Exercise:

Problem:

[link] shows two 7.50-cm-long glass slides illuminated by pure 589-nm wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on some debris at the other end, forming a wedge of air. How thick is the debris, if the dark bands are 1.00 mm apart?

Solution:

$4.42 \times 10^{-5}\,\text{m}$

Exercise:

Problem:

A soap bubble is 100 nm thick and illuminated by white light incident at a 45° angle to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

Exercise:

Problem:

An oil slick on water is 120 nm thick and illuminated by white light incident at a 45° angle to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

Solution:

for one phase change: 950 nm (infrared); for three phase changes: 317 nm (ultraviolet); Therefore, the oil film will appear black, since the reflected light is not in the visible part of the spectrum.

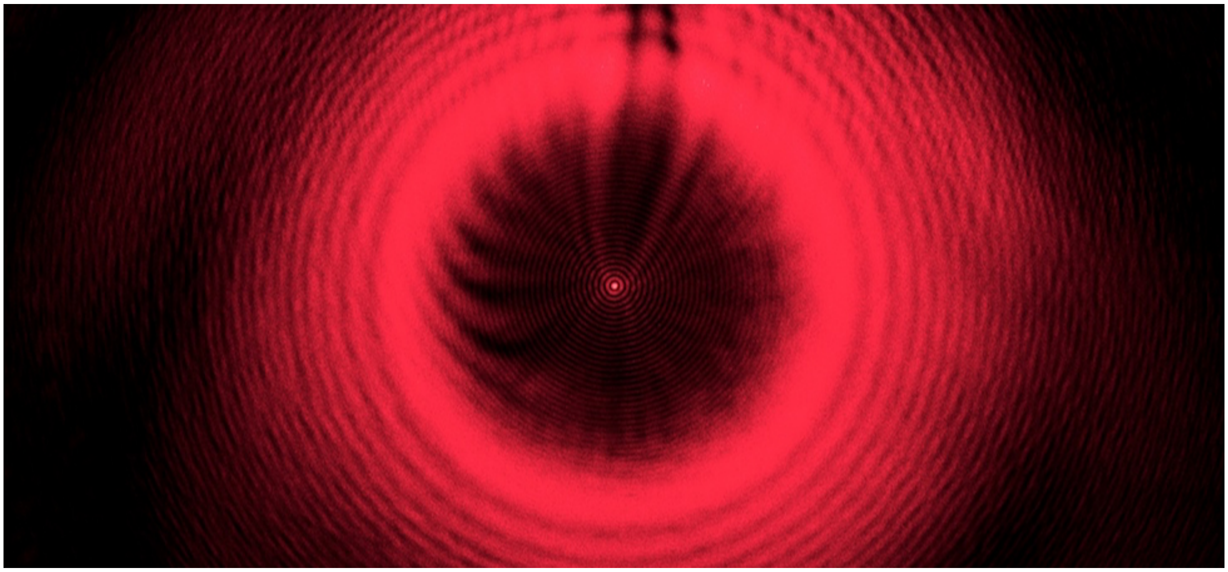
Glossary

interferometer	instrument that uses interference of waves to make measurements

Introduction class="introduction"

A steel ball bearing illuminated by a laser does not cast a sharp, circular shadow. Instead, a series of diffraction fringes and a central bright spot are observed. Known as Poisson’s spot, the effect was first predicted by Augustin-Jean Fresnel (1788–1827) as a consequence of diffraction of light waves. Based on principles of ray optics, Siméon-Denis Poisson (1781–1840) argued against Fresnel’s prediction. (credit: modification of work by Harvard Natural

Science Lecture
Demonstrations
)



Imagine passing a monochromatic light beam through a narrow opening—a slit just a little wider than the wavelength of the light. Instead of a simple shadow of the slit on the screen, you will see that an interference pattern appears, even though there is only one slit.

In the chapter on interference, we saw that you need two sources of waves for interference to occur. How can there be an interference pattern when we have only one slit? In The Nature of Light, we learned that, due to Huygens’s principle, we can imagine a wave front as equivalent to infinitely many point sources of waves. Thus, a wave from a slit can behave not as one wave but as an infinite number of point sources. These waves can interfere with each other, resulting in an interference pattern without the presence of a second slit. This phenomenon is called *diffraction*.

Another way to view this is to recognize that a slit has a small but finite width. In the preceding chapter, we implicitly regarded slits as objects with positions but no size. The widths of the slits were considered negligible. When the slits have finite widths, each point along the opening can be considered a point source of light—a foundation of Huygens’s principle.

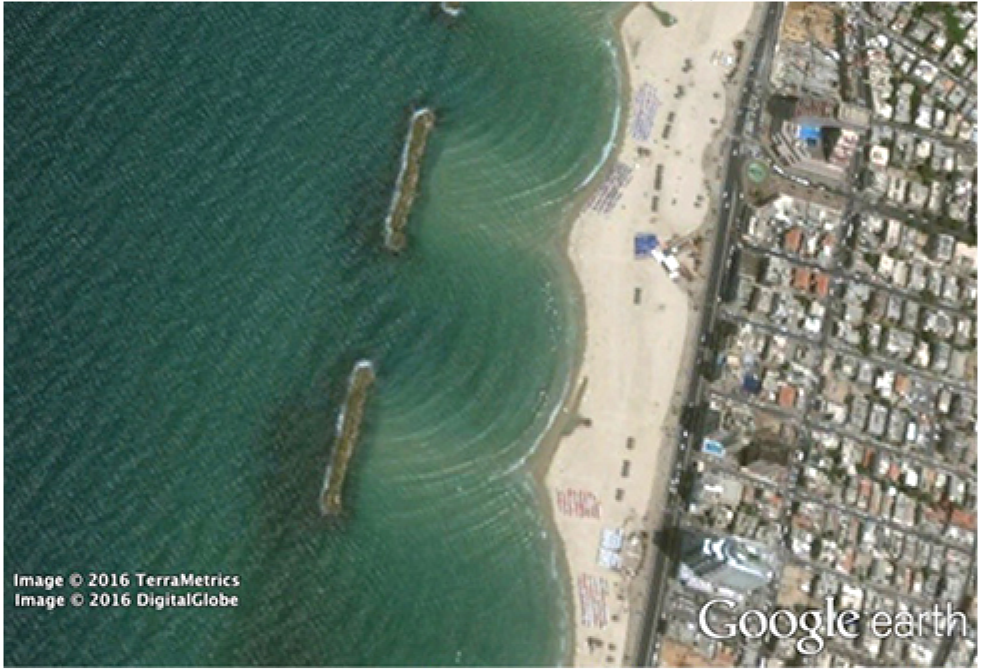
Because real-world optical instruments must have finite apertures (otherwise, no light can enter), diffraction plays a major role in the way we interpret the output of these optical instruments. For example, diffraction places limits on our ability to resolve images or objects. This is a problem that we will study later in this chapter.

Single-Slit Diffraction

By the end of this section, you will be able to:

- Explain the phenomenon of diffraction and the conditions under which it is observed
- Describe diffraction through a single slit

After passing through a narrow aperture (opening), a wave propagating in a specific direction tends to spread out. For example, sound waves that enter a room through an open door can be heard even if the listener is in a part of the room where the geometry of ray propagation dictates that there should only be silence. Similarly, ocean waves passing through an opening in a breakwater can spread throughout the bay inside. ([link]). The spreading and bending of sound and ocean waves are two examples of **diffraction**, which is the bending of a wave around the edges of an opening or an obstacle—a phenomenon exhibited by all types of waves.

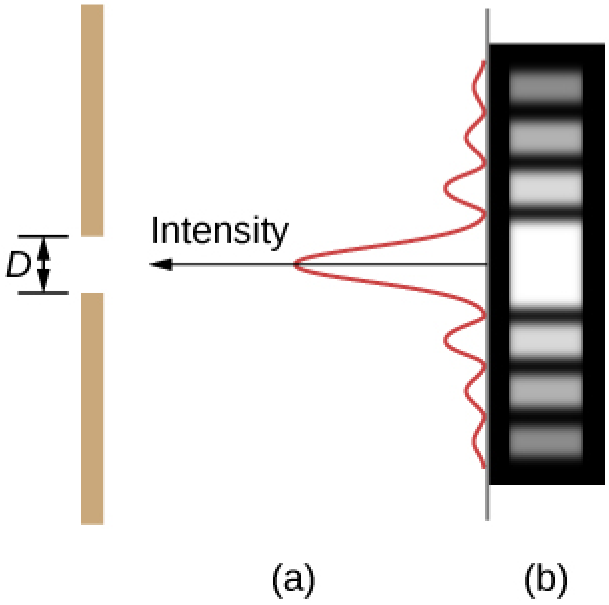


Because of the diffraction of waves, ocean waves entering through an opening in a breakwater can spread throughout the bay. (credit: modification of map data from Google Earth)

The diffraction of sound waves is apparent to us because wavelengths in the audible region are approximately the same size as the objects they encounter, a condition that must be satisfied if diffraction effects are to be observed easily. Since the wavelengths of visible light range from approximately 390 to 770 nm, most objects do not diffract light significantly. However, situations do occur in which apertures are small enough that the diffraction of light is observable. For example, if you place your middle and index fingers close together and look through the opening at a light bulb, you can see a rather clear diffraction pattern, consisting of light and dark lines running parallel to your fingers.

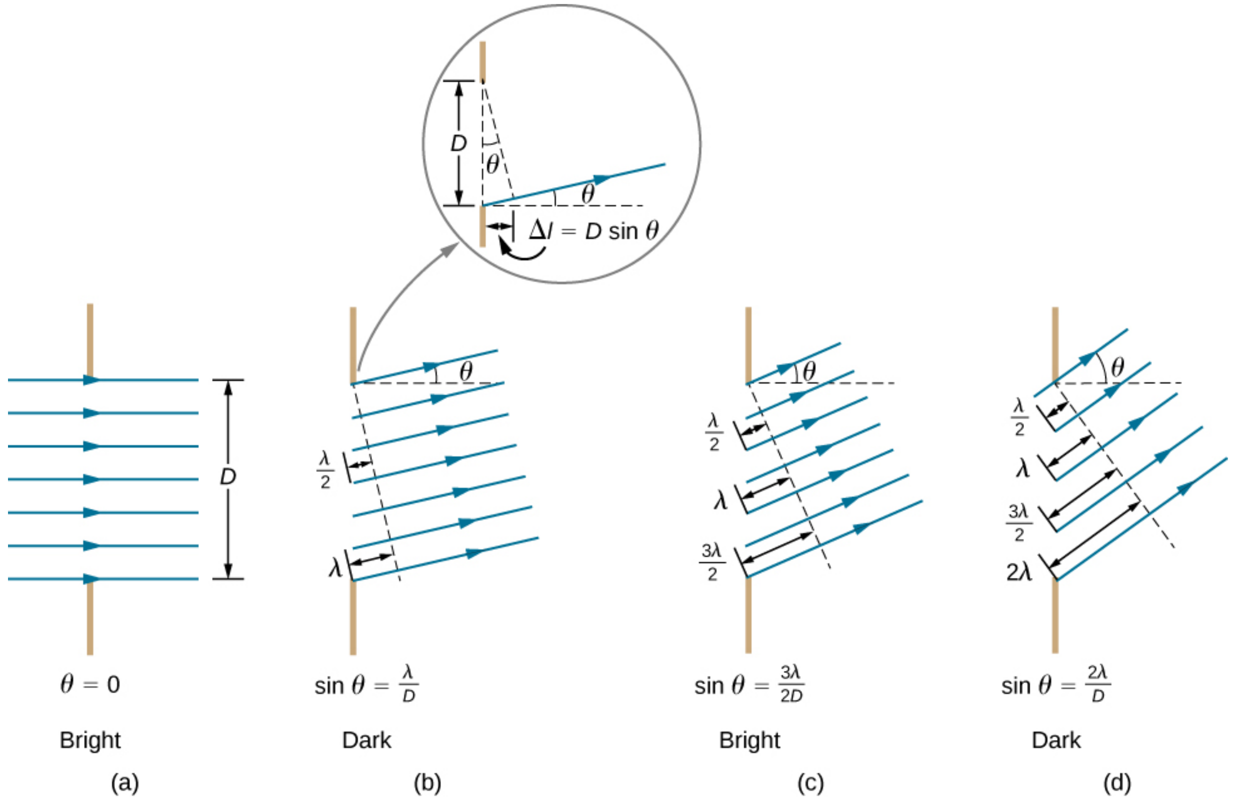
Diffraction through a Single Slit

Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings, which we discussed in the chapter on interference. [link] shows a single-slit diffraction pattern. Note that the central maximum is larger than maxima on either side and that the intensity decreases rapidly on either side. In contrast, a diffraction grating (Diffraction Gratings) produces evenly spaced lines that dim slowly on either side of the center.



Single-slit diffraction pattern. (a) Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The diagram shows the bright central maximum, and the dimmer and thinner maxima on either side.

The analysis of single-slit diffraction is illustrated in [link]. Here, the light arrives at the slit, illuminating it uniformly and is in phase across its width. We then consider light propagating onwards from different parts of the *same* slit. According to Huygens’s principle, every part of the wave front in the slit emits wavelets, as we discussed in The Nature of Light. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wave front of a wavelet.) Assuming the screen is very far away compared with the size of the slit, rays heading toward a common destination are nearly parallel. When they travel straight ahead, as in part (a) of the figure, they remain in phase, and we observe a central maximum. However, when rays travel at an angle θ relative to the original direction of the beam, each ray travels a different distance to a common location, and they can arrive in or out of phase. In part (b), the ray from the bottom travels a distance of one wavelength λ farther than the ray from the top. Thus, a ray from the center travels a distance $\lambda/2$ less than the one at the bottom edge of the slit, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom also cancel one another. In fact, each ray from the slit interferes destructively with another ray. In other words, a pair-wise cancellation of all rays results in a dark minimum in intensity at this angle. By symmetry, another minimum occurs at the same angle to the right of the incident direction (toward the bottom of the figure) of the light.



Light passing through a single slit is diffracted in all directions and may interfere constructively or destructively, depending on the angle. The difference in path length for rays from either side of the slit is seen to be $D \sin \theta$.

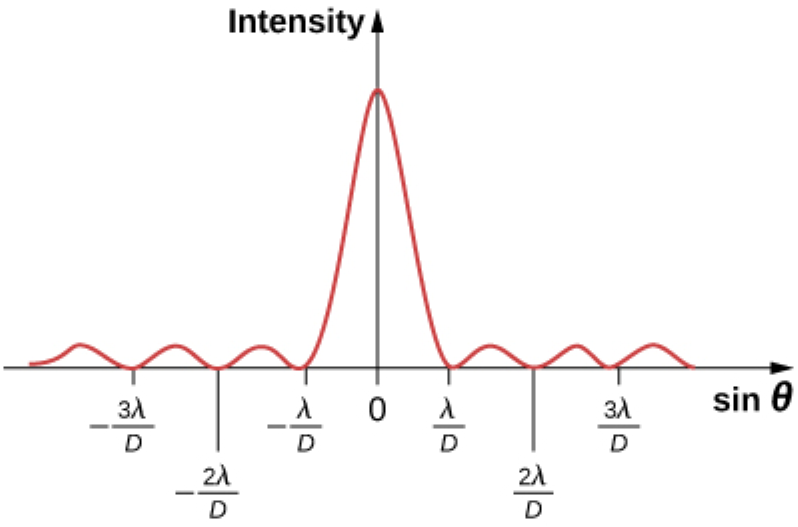
At the larger angle shown in part (c), the path lengths differ by $3\lambda/2$ for rays from the top and bottom of the slit. One ray travels a distance λ different from the ray from the bottom and arrives in phase, interfering constructively. Two rays, each from slightly above those two, also add constructively. Most rays from the slit have another ray to interfere with constructively, and a maximum in intensity occurs at this angle. However, not all rays interfere constructively for this situation, so the maximum is not as intense as the central maximum. Finally, in part (d), the angle shown is large enough to produce a second minimum. As seen in the figure, the difference in path length for rays from either side of the slit is $D \sin \theta$, and we see that a destructive minimum is obtained when this distance is an integral multiple of the wavelength.

Thus, to obtain **destructive interference for a single slit**,

Note:
Equation:

$$D\sin\theta = m\lambda,\ \text{for}\ m = \pm1, \pm2, \pm3, \ldots(\text{destructive}),$$

where D is the slit width, λ is the light’s wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. [link] shows a graph of intensity for single-slit interference, and it is apparent that the maxima on either side of the central maximum are much less intense and not as wide. This effect is explored in Double-Slit Diffraction.

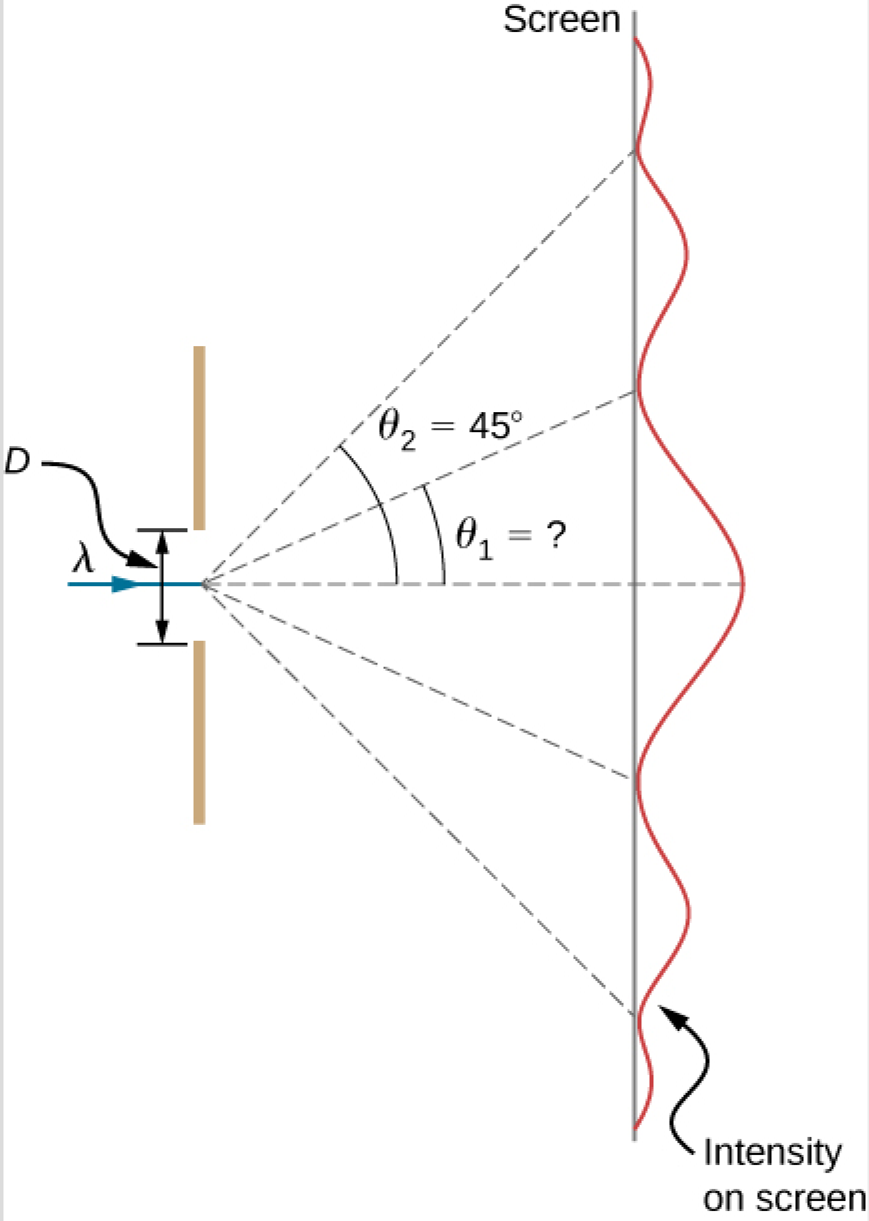


A graph of single-slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact, the central maximum is six times higher than shown here.

Example:

Calculating Single-Slit Diffraction

Visible light of wavelength 550 nm falls on a single slit and produces its second diffraction minimum at an angle of 45.0° relative to the incident direction of the light, as in [link]. (a) What is the width of the slit? (b) At what angle is the first minimum produced?


Screen
$\theta_2 = 45°$
D
λ
$\theta_1 = ?$
Intensity on screen

In this example, we analyze a graph of the single-slit diffraction pattern.

Strategy

From the given information, and assuming the screen is far away from the slit, we can use the equation $D \sin \theta = m\lambda$ first to find D, and again to find the angle for the first minimum θ_1.

Solution

a. We are given that $\lambda = 550\,\text{nm}$, $m = 2$, and $\theta_2 = 45.0^\circ$. Solving the equation $D \sin \theta = m\lambda$ for D and substituting known values gives
Equation:

$$D = \frac{m\lambda}{\sin \theta_2} = \frac{2(550\,\text{nm})}{\sin 45.0^\circ} = \frac{1100 \times 10^{-9}\,\text{m}}{0.707} = 1.56 \times 10^{-6}\,\text{m}.$$

b. Solving the equation $D \sin \theta = m\lambda$ for $\sin \theta_1$ and substituting the known values gives
Equation:

$$\sin \theta_1 = \frac{m\lambda}{D} = \frac{1(550 \times 10^{-9}\,\text{m})}{1.56 \times 10^{-6}\,\text{m}}.$$

Thus the angle θ_1 is
Equation:

$$\theta_1 = \sin^{-1} 0.354 = 20.7^\circ.$$

Significance

We see that the slit is narrow (it is only a few times greater than the wavelength of light). This is consistent with the fact that light must interact with an object comparable in size to its wavelength in order to exhibit significant wave effects such as this single-slit diffraction pattern. We also see that the central maximum extends 20.7° on either side of the original beam, for a width of about 41°. The angle between the first and second minima is only about $24^\circ(45.0^\circ - 20.7^\circ)$. Thus, the second maximum is only about half as wide as the central maximum.

Note:

Exercise:

Problem:

Check Your Understanding Suppose the slit width in [link] is increased to $1.8 \times 10^{-6}\,\text{m}$. What are the new angular positions for the first, second, and third minima? Would a fourth minimum exist?

Solution:

17.8°, 37.7°, 66.4°; no

Summary

- Diffraction can send a wave around the edges of an opening or other obstacle.
- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.

Conceptual Questions

Exercise:

Problem:

As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

Solution:

The diffraction pattern becomes wider.

Exercise:

Problem: Compare interference and diffraction.

Exercise:

Problem:

If you and a friend are on opposite sides of a hill, you can communicate with walkie-talkies but not with flashlights. Explain.

Solution:

Walkie-talkies use radio waves whose wavelengths are comparable to the size of the hill and are thus able to diffract around the hill. Visible wavelengths of the flashlight travel as rays at this size scale.

Exercise:

Problem:

What happens to the diffraction pattern of a single slit when the entire optical apparatus is immersed in water?

Exercise:

Problem:

In our study of diffraction by a single slit, we assume that the length of the slit is much larger than the width. What happens to the diffraction pattern if these two dimensions were comparable?

Solution:

The diffraction pattern becomes two-dimensional, with main fringes, which are now spots, running in perpendicular directions and fainter spots in intermediate directions.

Exercise:

Problem:

A rectangular slit is twice as wide as it is high. Is the central diffraction peak wider in the vertical direction or in the horizontal direction?

Problems

Exercise:

Problem:

(a) At what angle is the first minimum for 550-nm light falling on a single slit of width $1.00 \mu\text{m}$? (b) Will there be a second minimum?

Solution:

a. 33.4°; b. no

Exercise:

Problem:

(a) Calculate the angle at which a 2.00-μm-wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?

Exercise:

Problem:

(a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of 28.0° ? (b) At what angle will the second minimum be?

Solution:

a. $1.35 \times 10^{-6}\,\text{m}$; b. 69.9°

Exercise:

Problem:

(a) What is the width of a single slit that produces its first minimum at 60.0° for 600-nm light? (b) Find the wavelength of light that has its first minimum at 62.0°.

Exercise:

Problem:

Find the wavelength of light that has its third minimum at an angle of 48.6° when it falls on a single slit of width $3.00\mu\text{m}$.

Solution:

750 nm

Exercise:

Problem:

(a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width $7.50\mu\text{m}$. At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?

Exercise:

Problem:

Consider a single-slit diffraction pattern for $\lambda = 589\text{ nm}$, projected on a screen that is 1.00 m from a slit of width 0.25 mm. How far from the center of the pattern are the centers of the first and second dark fringes?

Solution:

2.4 mm, 4.7 mm

Exercise:

Problem:

(a) Find the angle between the first minima for the two sodium vapor lines, which have wavelengths of 589.1 and 589.6 nm, when they fall upon a single slit of width $2.00 \mu m$. (b) What is the distance between these minima if the diffraction pattern falls on a screen 1.00 m from the slit? (c) Discuss the ease or difficulty of measuring such a distance.

Exercise:

Problem:

(a) What is the minimum width of a single slit (in multiples of λ) that will produce a first minimum for a wavelength λ ? (b) What is its minimum width if it produces 50 minima? (c) 1000 minima?

Solution:

a. 1.00λ; b. 50.0λ; c. 1000λ

Exercise:

Problem:

(a) If a single slit produces a first minimum at 14.5°, at what angle is the second-order minimum? (b) What is the angle of the third-order minimum? (c) Is there a fourth-order minimum? (d) Use your answers to illustrate how the angular width of the central maximum is about twice the angular width of the next maximum (which is the angle between the first and second minima).

Exercise:

Problem:

If the separation between the first and the second minima of a single-slit diffraction pattern is 6.0 mm, what is the distance between the screen and the slit? The light wavelength is 500 nm and the slit width is 0.16 mm.

Solution:

1.92 m

Exercise:

Problem:

A water break at the entrance to a harbor consists of a rock barrier with a 50.0-m-wide opening. Ocean waves of 20.0-m wavelength approach the opening straight on. At what angles to the incident direction are the boats inside the harbor most protected against wave action?

Exercise:

Problem:

An aircraft maintenance technician walks past a tall hangar door that acts like a single slit for sound entering the hangar. Outside the door, on a line perpendicular to the opening in the door, a jet engine makes a 600-Hz sound. At what angle with the door will the technician observe the first minimum in sound intensity if the vertical opening is 0.800 m wide and the speed of sound is 340 m/s?

Solution:

45.1°

Glossary

destructive interference for a single slit	occurs when the width of the slit is comparable to the wavelength of light illuminating it
diffraction	bending of a wave around the edges of an opening or an obstacle

Intensity in Single-Slit Diffraction

By the end of this section, you will be able to:

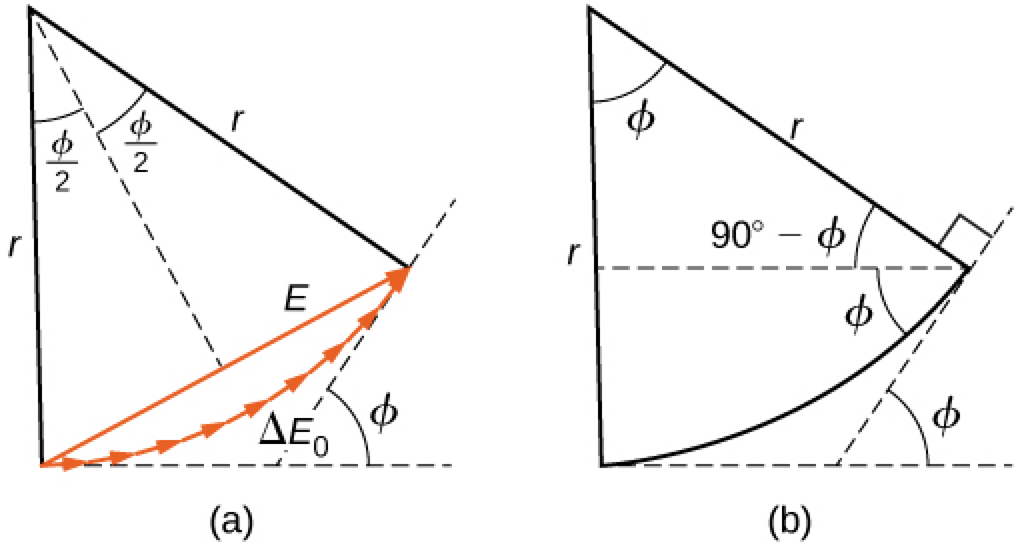
- Calculate the intensity relative to the central maximum of the single-slit diffraction peaks
- Calculate the intensity relative to the central maximum of an arbitrary point on the screen

To calculate the intensity of the diffraction pattern, we follow the phasor method used for calculations with ac circuits in Alternating-Current Circuits. If we consider that there are *N* Huygens sources across the slit shown in [link], with each source separated by a distance *D*/*N* from its adjacent neighbors, the path difference between waves from adjacent sources reaching the arbitrary point *P* on the screen is $(D/N)\sin\theta$. This distance is equivalent to a phase difference of $(2\pi D/\lambda N)\sin\theta$. The phasor diagram for the waves arriving at the point whose angular position is θ is shown in [link]. The amplitude of the phasor for each Huygens wavelet is ΔE_0, the amplitude of the resultant phasor is *E*, and the phase difference between the wavelets from the first and the last sources is

Equation:

$$\phi = \left(\frac{2\pi}{\lambda}\right) D\sin\theta.$$

With $N \to \infty$, the phasor diagram approaches a circular arc of length $N\Delta E_0$ and radius *r*. Since the length of the arc is $N\Delta E_0$ for any ϕ, the radius *r* of the arc must decrease as ϕ increases (or equivalently, as the phasors form tighter spirals).



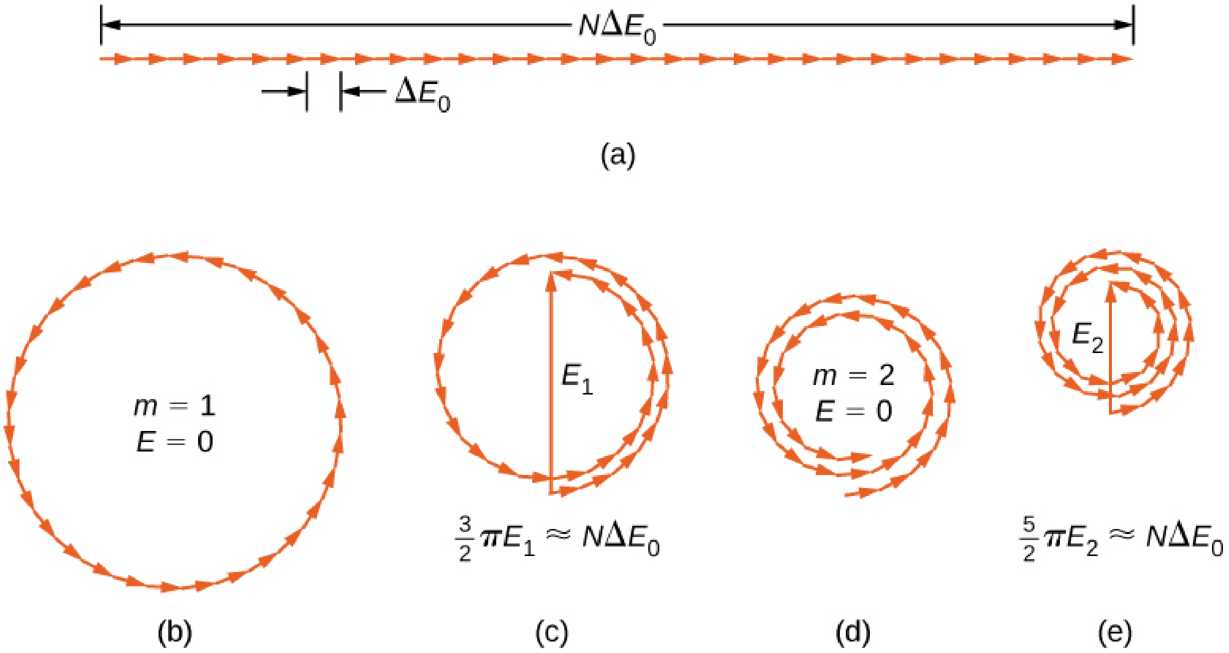
(a) Phasor diagram corresponding to the angular position θ in the single-slit diffraction pattern. The phase difference between the wavelets from the first and last sources is $\phi = (2\pi/\lambda)D\sin\theta$. (b) The geometry of the phasor diagram.

The phasor diagram for $\phi = 0$ (the center of the diffraction pattern) is shown in [link](a) using $N = 30$. In this case, the phasors are laid end to end in a straight line of length $N\Delta E_0$, the radius *r* goes to infinity, and the resultant has its maximum value $E = N\Delta E_0$. The intensity of the light can be obtained using the relation $I = \frac{1}{2}c\varepsilon_0 E^2$ from Electromagnetic Waves. The intensity of the maximum is then

Equation:

$$I_0 = \frac{1}{2}c\varepsilon_0(N\Delta E_0)^2 = \frac{1}{2\mu_0 c}(N\Delta E_0)^2,$$

where $\varepsilon_0 = 1/\mu_0 c^2$. The phasor diagrams for the first two zeros of the diffraction pattern are shown in parts (b) and (d) of the figure. In both cases, the phasors add to zero, after rotating through $\phi = 2\pi$ rad for $m = 1$ and 4π rad for $m = 2$.


$N\Delta E_0$
ΔE_0
(a)
$m = 1$
$E = 0$
E_1
$\frac{3}{2}\pi E_1 \approx N\Delta E_0$
$m = 2$
$E = 0$
E_2
$\frac{5}{2}\pi E_2 \approx N\Delta E_0$
(b)
(c)
(d)
(e)

Phasor diagrams (with 30 phasors) for various points on the single-slit diffraction pattern. Multiple rotations around a given circle have been separated slightly so that the phasors can be seen. (a) Central maximum, (b) first minimum, (c) first maximum beyond central maximum, (d) second minimum, and (e) second maximum beyond central maximum.

The next two maxima beyond the central maxima are represented by the phasor diagrams of parts (c) and (e). In part (c), the phasors have rotated through $\phi = 3\pi$ rad and have formed a resultant phasor of magnitude E_1. The length of the arc formed by the phasors is $N\Delta E_0$. Since this corresponds to 1.5 rotations around a circle of diameter E_1, we have

Equation:

$$\frac{3}{2}\pi E_1 \approx N\Delta E_0,$$

so

Equation:

$$E_1 = \frac{2N\Delta E_0}{3\pi}$$

and
Equation:

$$I_1 = \frac{1}{2\mu_0 c} E_1^2 = \frac{4(N\Delta E_0)^2}{(9\pi^2)(2\mu_0 c)} \approx 0.045 I_0,$$

where
Equation:

$$I_0 = \frac{(N\Delta E_0)^2}{2\mu_0 c}.$$

In part (e), the phasors have rotated through $\phi = 5\pi$ rad, corresponding to 2.5 rotations around a circle of diameter E_2 and arc length $N\Delta E_0$. This results in $I_2 \approx 0.016 I_0$. The proof is left as an exercise for the student ([link]).

These two maxima actually correspond to values of ϕ slightly less than 3π rad and 5π rad. Since the total length of the arc of the phasor diagram is always $N\Delta E_0$, the radius of the arc decreases as ϕ increases. As a result, E_1 and E_2 turn out to be slightly larger for arcs that have not quite curled through 3π rad and 5π rad, respectively. The exact values of ϕ for the maxima are investigated in [link]. In solving that problem, you will find that they are less than, but very close to, $\phi = 3\pi, 5\pi, 7\pi, \ldots$ rad.

To calculate the intensity at an arbitrary point *P* on the screen, we return to the phasor diagram of [link]. Since the arc subtends an angle ϕ at the center of the circle,
Equation:

$$N\Delta E_0 = r\phi$$

and
Equation:

$$\sin\left(\frac{\phi}{2}\right) = \frac{E}{2r}.$$

where E is the amplitude of the resultant field. Solving the second equation for E and then substituting r from the first equation, we find
Equation:

$$E = 2r\sin\frac{\phi}{2} = 2\frac{N\Delta E_o}{\phi}\sin\frac{\phi}{2}.$$

Now defining

Note:
Equation:

$$\beta = \frac{\phi}{2} = \frac{\pi D\sin\theta}{\lambda}$$

we obtain

Note:
Equation:

$$E = N\Delta E_0\frac{\sin\beta}{\beta}$$

This equation relates the amplitude of the resultant field at any point in the diffraction pattern to the amplitude $N\Delta E_0$ at the central maximum. The intensity is proportional to the square of the amplitude, so

Note:
Equation:

$$I = I_0 \left(\frac{\sin \beta}{\beta} \right)^2$$

where $I_0 = (N\Delta E_0)^2 / 2\mu_0 c$ is the intensity at the center of the pattern.

For the central maximum, $\phi = 0$, β is also zero and we see from l'Hôpital's rule that $\lim_{\beta \to 0} (\sin \beta / \beta) = 1$, so that $\lim_{\phi \to 0} I = I_0$. For the next maximum, $\phi = 3\pi$ rad, we have $\beta = 3\pi/2$ rad and when substituted into [link], it yields
Equation:

$$I_1 = I_0 \left(\frac{\sin 3\pi/2}{3\pi/2} \right)^2 \approx 0.045 I_0,$$

in agreement with what we found earlier in this section using the diameters and circumferences of phasor diagrams. Substituting $\phi = 5\pi$ rad into [link] yields a similar result for I_2.

A plot of [link] is shown in [link] and directly below it is a photograph of an actual diffraction pattern. Notice that the central peak is much brighter than the others, and that the zeros of the pattern are located at those points where $\sin \beta = 0$, which occurs when $\beta = m\pi$ rad. This corresponds to
Equation:

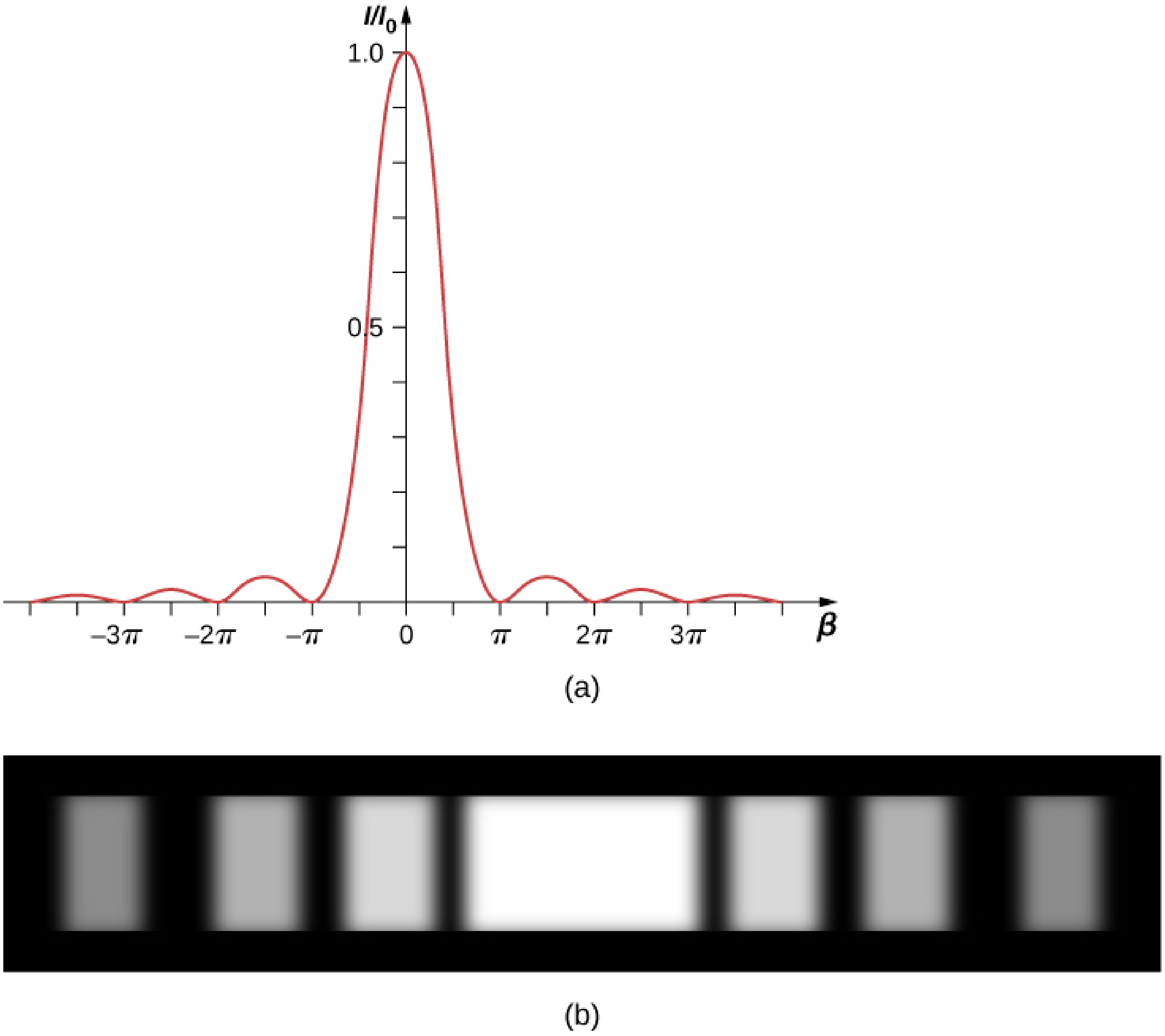
$$\frac{\pi D\sin\theta}{\lambda}=m\pi,$$

or

Equation:

$$D\sin\theta=m\lambda,$$

which is [link].



(a) The calculated intensity distribution of a single-slit diffraction pattern. (b) The actual diffraction pattern.

Example:

Intensity in Single-Slit Diffraction

Light of wavelength 550 nm passes through a slit of width $2.00\ \mu\text{m}$ and produces a diffraction pattern similar to that shown in [link]. (a) Find the locations of the first two minima in terms of the angle from the central maximum and (b) determine the intensity relative to the central maximum at a point halfway between these two minima.

Strategy

The minima are given by [link], $D \sin \theta = m\lambda$. The first two minima are for $m = 1$ and $m = 2$. [link] and [link] can be used to determine the intensity once the angle has been worked out.

Solution

a. Solving [link] for θ gives us $\theta_m = \sin^{-1}(m\lambda/D)$, so that

   Equation:

$$\theta_1 = \sin^{-1}\left(\frac{(+1)\left(550\ \times\ 10^{-9}\ \text{m}\right)}{2.00\ \times\ 10^{-6}\ \text{m}}\right) = +16.0^\circ$$

   and

   Equation:

$$\theta_2 = \sin^{-1}\left(\frac{(+2)\left(550\ \times\ 10^{-9}\ \text{m}\right)}{2.00\ \times\ 10^{-6}\ \text{m}}\right) = +33.4^\circ.$$

b. The halfway point between θ_1 and θ_2 is

   Equation:

$$\theta = (\theta_1 + \theta_2)/2 = (16.0^\circ + 33.4^\circ)/2 = 24.7^\circ.$$

[link] gives

Equation:

$$\beta = \frac{\pi D \sin \theta}{\lambda} = \frac{\pi \left(2.00 \ \times \ 10^{-6} \, \text{m}\right) \sin \left(24.7^{\circ}\right)}{\left(550 \ \times \ 10^{-9} \, \text{m}\right)} = 1.52\pi \, \text{or} \, 4.77 \, \text{rad.}$$

From [link], we can calculate

Equation:

$$\frac{I}{I_o} = \left(\frac{\sin \beta}{\beta}\right)^2 = \left(\frac{\sin \ (4.77)}{4.77}\right)^2 = \left(\frac{-0.9985}{4.77}\right)^2 = 0.044.$$

Significance

This position, halfway between two minima, is very close to the location of the maximum, expected near $\beta = 3\pi/2,$ or 1.5π.

Note:

Exercise:

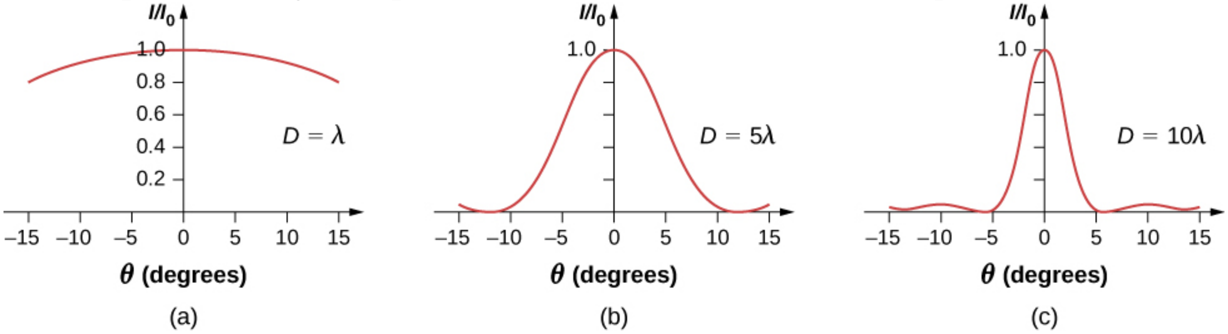
Problem:

Check Your Understanding For the experiment in [link], at what angle from the center is the third maximum and what is its intensity relative to the central maximum?

Solution:

74.3°, $0.0083 I_0$

If the slit width D is varied, the intensity distribution changes, as illustrated in [link]. The central peak is distributed over the region from $\sin \theta = -\lambda/D$ to $\sin \theta = +\lambda/D$. For small θ, this corresponds to an angular width $\Delta\theta \approx 2\lambda/D$. Hence, an increase in the slit width results in a decrease in the **width of the central peak**. For a slit with $D \gg \lambda$, the central peak is very sharp, whereas if $D \approx \lambda$, it becomes quite broad.



Single-slit diffraction patterns for various slit widths. As the slit width D increases from $D = \lambda$ to 5λ and then to 10λ, the width of the central peak decreases as the angles for the first minima decrease as predicted by [link].

Note:
A diffraction experiment in optics can require a lot of preparation but this simulation by Andrew Duffy offers not only a quick set up but also the ability to change the slit width instantly. Run the simulation and select “Single slit.” You can adjust the slit width and see the effect on the diffraction pattern on a screen and as a graph.

Summary

- The intensity pattern for diffraction due to a single slit can be calculated using phasors as
  Equation:

$$I = I_0 \left(\frac{\sin \beta}{\beta} \right)^2,$$

where $\beta = \frac{\phi}{2} = \frac{\pi D \sin\theta}{\lambda}$, D is the slit width, λ is the wavelength, and θ is the angle from the central peak.

Conceptual Questions

Exercise:

Problem:

In [link], the parameter β looks like an angle but is not an angle that you can measure with a protractor in the physical world. Explain what β represents.

Solution:

The parameter $\beta = \phi/2$ is the arc angle shown in the phasor diagram in [link]. The phase difference between the first and last Huygens wavelet across the single slit is 2β and is related to the curvature of the arc that forms the resultant phasor that determines the light intensity.

Problems

Exercise:

Problem:

A single slit of width $3.0\ \mu$m is illuminated by a sodium yellow light of wavelength 589 nm. Find the intensity at a 15° angle to the axis in terms of the intensity of the central maximum.

Exercise:

Problem:

A single slit of width 0.1 mm is illuminated by a mercury light of wavelength 576 nm. Find the intensity at a 10° angle to the axis in terms of the intensity of the central maximum.

Solution:

$I/I_0 = 2.2 \times 10^{-5}$

Exercise:

Problem:

The width of the central peak in a single-slit diffraction pattern is 5.0 mm. The wavelength of the light is 600 nm, and the screen is 2.0 m from the slit. (a) What is the width of the slit? (b) Determine the ratio of the intensity at 4.5 mm from the center of the pattern to the intensity at the center.

Exercise:

Problem:

Consider the single-slit diffraction pattern for $\lambda = 600\text{ nm}$, $D = 0.025\text{ mm}$, and $x = 2.0\text{ m}$. Find the intensity in terms of I_o at $\theta = 0.5^\circ$, 1.0°, 1.5°, 3.0°, and 10.0°.

Solution:

$0.63I_0,\ 0.11I_0,\ 0.0067I_0,\ 0.0062I_0,\ 0.00088I_0$

Glossary

width of the central peak
angle between the minimum for $m = 1$ and $m = -1$

Double-Slit Diffraction

By the end of this section, you will be able to:

- Describe the combined effect of interference and diffraction with two slits, each with finite width
- Determine the relative intensities of interference fringes within a diffraction pattern
- Identify missing orders, if any

When we studied interference in Young’s double-slit experiment, we ignored the diffraction effect in each slit. We assumed that the slits were so narrow that on the screen you saw only the interference of light from just two point sources. If the slit is smaller than the wavelength, then [link](a) shows that there is just a spreading of light and no peaks or troughs on the screen. Therefore, it was reasonable to leave out the diffraction effect in that chapter. However, if you make the slit wider, [link](b) and (c) show that you cannot ignore diffraction. In this section, we study the complications to the double-slit experiment that arise when you also need to take into account the diffraction effect of each slit.

To calculate the diffraction pattern for two (or any number of) slits, we need to generalize the method we just used for a single slit. That is, across each slit, we place a uniform distribution of point sources that radiate Huygens wavelets, and then we sum the wavelets from all the slits. This gives the intensity at any point on the screen. Although the details of that calculation can be complicated, the final result is quite simple:

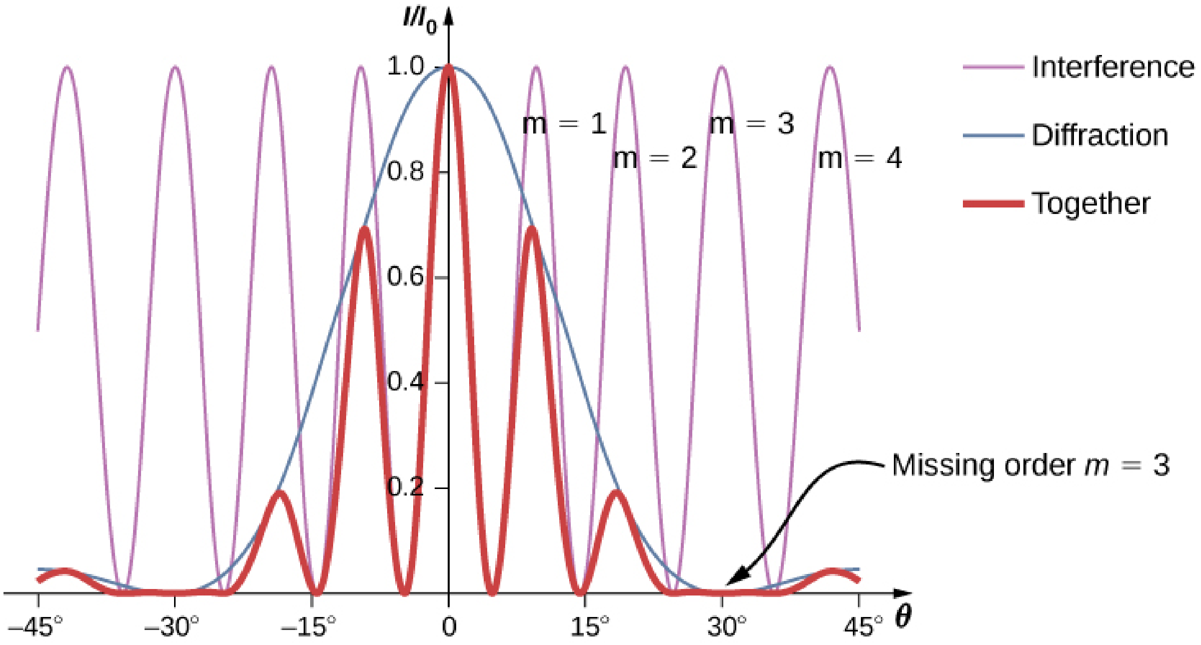
Note:

Two-Slit Diffraction Pattern

The diffraction pattern of two slits of width D that are separated by a distance d is the interference pattern of two point sources separated by d multiplied by the diffraction pattern of a slit of width D.

In other words, the *locations* of the interference fringes are given by the equation $d\sin\theta = m\lambda$, the same as when we considered the slits to be point sources, but the *intensities* of the fringes are now reduced by diffraction effects, according to [link]. [Note that in the chapter on interference, we wrote $d\sin\theta = m\lambda$ and used the integer m to refer to interference fringes. [link] also uses m, but this time to refer to diffraction minima. If both equations are used simultaneously, it is good practice to use a different variable (such as n) for one of these integers in order to keep them distinct.]

Interference and diffraction effects operate simultaneously and generally produce minima at different angles. This gives rise to a complicated pattern on the screen, in which some of the maxima of interference from the two slits are missing if the maximum of the interference is in the same direction as the minimum of the diffraction. We refer to such a missing peak as a **missing order**. One example of a diffraction pattern on the screen is shown in [link]. The solid line with multiple peaks of various heights is the intensity observed on the screen. It is a product of the interference pattern of waves from separate slits and the diffraction of waves from within one slit.



Diffraction from a double slit. The purple line with peaks of the same height are from the interference of the waves from two slits; the blue line with one big hump in the middle is the diffraction of waves from within one slit; and the thick red line is the product of the two, which is the pattern observed on the screen. The plot shows the expected result for a slit width $D = 2\lambda$ and slit separation $d = 6\lambda$. The maximum of $m = \pm 3$ order for the interference is missing because the minimum of the diffraction occurs in the same direction.

Example:

Intensity of the Fringes

[link] shows that the intensity of the fringe for $m = 3$ is zero, but what about the other fringes? Calculate the intensity for the fringe at $m = 1$ relative to I_0, the intensity of the central peak.

Strategy

Determine the angle for the double-slit interference fringe, using the equation from Interference, then determine the relative intensity in that direction due to diffraction by using [link].

Solution

From the chapter on interference, we know that the bright interference fringes occur at $d \sin \theta = m\lambda$, or

Equation:

$$\sin\theta = \frac{m\lambda}{d}.$$

From [link],

Equation:

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2, \text{ where } \beta = \frac{\phi}{2} = \frac{\pi D\sin\theta}{\lambda}.$$

Substituting from above,

Equation:

$$\beta = \frac{\pi D\sin\theta}{\lambda} = \frac{\pi D}{\lambda}\cdot\frac{m\lambda}{d} = \frac{m\pi D}{d}.$$

For $D = 2\lambda$, $d = 6\lambda$, and $m = 1$,

Equation:

$$\beta = \frac{(1)\pi(2\lambda)}{(6\lambda)} = \frac{\pi}{3}.$$

Then, the intensity is

Equation:

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2 = I_0\left(\frac{\sin\ (\pi/3)}{\pi/3}\right)^2 = 0.684 I_0.$$

Significance

Note that this approach is relatively straightforward and gives a result that is almost exactly the same as the more complicated analysis using phasors to work out the intensity values of the double-slit interference (thin line in [link]). The phasor approach accounts for the downward slope in the diffraction intensity (blue line) so that the peak *near* $m = 1$ occurs at a value of θ ever so slightly smaller than we have shown here.

Example:

Two-Slit Diffraction

Suppose that in Young’s experiment, slits of width 0.020 mm are separated by 0.20 mm. If the slits are illuminated by monochromatic light of wavelength 500 nm, how many bright fringes are observed in the central peak of the diffraction pattern?

Solution

From [link], the angular position of the first diffraction minimum is $\theta \approx \sin\theta = \frac{\lambda}{D} = \frac{5.0 \times 10^{-7}\,\text{m}}{2.0 \times 10^{-5}\,\text{m}} = 2.5 \times 10^{-2}\,\text{rad}$.
Using $d\sin\theta = m\lambda$ for $\theta = 2.5 \times 10^{-2}\,\text{rad}$, we find

Equation:

$$m = \frac{d\sin\theta}{\lambda} = \frac{(0.20\,\text{mm})\ 2.5 \times 10^{-2}\,\text{rad}}{5.0 \times 10^{-7}\,\text{m}} = 10,$$

which is the maximum interference order that fits inside the central peak. We note that $m = \pm 10$ are missing orders as θ matches exactly. Accordingly, we observe bright fringes for

Equation:

$$m = -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5, +6, +7, +8, \text{ and } +9$$

for a total of 19 bright fringes.

Note:

Exercise:

Problem:

Check Your Understanding For the experiment in [link], show that $m = 20$ is also a missing order.

Solution:

From $d\sin\theta = m\lambda$, the interference maximum occurs at 2.87° for $m = 20$. From [link], this is also the angle for the second diffraction minimum. (*Note:* Both equations use the index *m* but they refer to separate phenomena.)

Note:

Explore the effects of double-slit diffraction. In this simulation written by Fu-Kwun Hwang, select $N = 2$ using the slider and see what happens when you control the slit width, slit separation and the wavelength. Can you make an order go “missing?”

Summary

- With real slits with finite widths, the effects of interference and diffraction operate simultaneously to form a complicated intensity pattern.

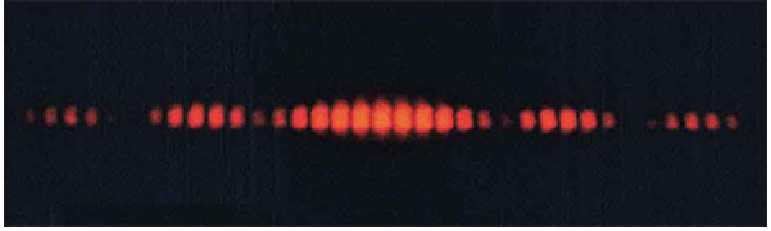
- Relative intensities of interference fringes within a diffraction pattern can be determined.
- Missing orders occur when an interference maximum and a diffraction minimum are located together.

Conceptual Questions

Exercise:

Problem:

Shown below is the central part of the interference pattern for a pure wavelength of red light projected onto a double slit. The pattern is actually a combination of single- and double-slit interference. Note that the bright spots are evenly spaced. Is this a double- or single-slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single- or double-slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.



(credit: PASCO)

Problems

Exercise:

Problem:

Two slits of width $2\ \mu\mathrm{m}$, each in an opaque material, are separated by a center-to-center distance of $6\ \mu\mathrm{m}$. A monochromatic light of wavelength 450 nm is incident on the double-slit. One finds a combined interference and diffraction pattern on the screen.

(a) How many peaks of the interference will be observed in the central maximum of the diffraction pattern?

(b) How many peaks of the interference will be observed if the slit width is doubled while keeping the distance between the slits same?

(c) How many peaks of interference will be observed if the slits are separated by twice the distance, that is, $12\ \mu\mathrm{m}$, while keeping the widths of the slits same?

(d) What will happen in (a) if instead of 450-nm light another light of wavelength 680 nm is used?

(e) What is the value of the ratio of the intensity of the central peak to the intensity of the next bright peak in (a)?

(f) Does this ratio depend on the wavelength of the light?

(g) Does this ratio depend on the width or separation of the slits?

Exercise:

Problem:

A double slit produces a diffraction pattern that is a combination of single- and double-slit interference. Find the ratio of the width of the slits to the separation between them, if the first minimum of the single-slit pattern falls on the fifth maximum of the double-slit pattern. (This will greatly reduce the intensity of the fifth maximum.)

Solution:

0.200

Exercise:

Problem:

For a double-slit configuration where the slit separation is four times the slit width, how many interference fringes lie in the central peak of the diffraction pattern?

Exercise:

Problem:

Light of wavelength 500 nm falls normally on 50 slits that are $2.5 \times 10^{-3}\,\mathrm{mm}$ wide and spaced $5.0 \times 10^{-3}\,\mathrm{mm}$ apart. How many interference fringes lie in the central peak of the diffraction pattern?

Solution:

3

Exercise:

Problem:

A monochromatic light of wavelength 589 nm incident on a double slit with slit width $2.5\,\mu\mathrm{m}$ and unknown separation results in a diffraction pattern containing nine interference peaks inside the central maximum. Find the separation of the slits.

Exercise:

Problem:

When a monochromatic light of wavelength 430 nm incident on a double slit of slit separation $5\ \mu\text{m}$, there are 11 interference fringes in its central maximum. How many interference fringes will be in the central maximum of a light of the same wavelength and slit widths, but a new slit separation of $4\ \mu\text{m}$?

Solution:

9

Exercise:

Problem:

Determine the intensities of two interference peaks other than the central peak in the central maximum of the diffraction, if possible, when a light of wavelength 628 nm is incident on a double slit of width 500 nm and separation 1500 nm. Use the intensity of the central spot to be $1\ \text{mW}/\text{cm}^2$.

Glossary

missing order	interference maximum that is not seen because it coincides with a diffraction minimum
two-slit diffraction pattern	diffraction pattern of two slits of width *D* that are separated by a distance *d* is the interference pattern of two point sources separated by *d* multiplied by the diffraction pattern of a slit of width *D*

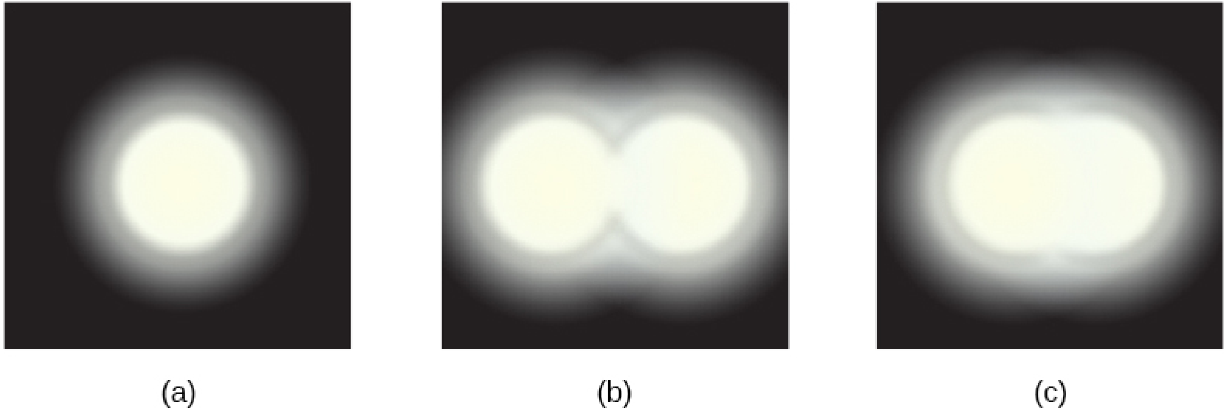
Circular Apertures and Resolution

By the end of this section, you will be able to:

- Describe the diffraction limit on resolution
- Describe the diffraction limit on beam propagation

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. This can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—but diffraction also limits the detail we can obtain in images.

[link](a) shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, we obtain a spot with a fuzzy edge surrounded by circles of light. This pattern is caused by diffraction, similar to that produced by a single slit. Light from different parts of the circular aperture interferes constructively and destructively. The effect is most noticeable when the aperture is small, but the effect is there for large apertures as well.



(a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point-light sources that are close to one another produce overlapping images because of diffraction. (c) If the sources are closer together, they cannot be distinguished or resolved.

How does diffraction affect the detail that can be observed when light passes through an aperture? [link](b) shows the diffraction pattern produced by two point-light sources that are close to one another. The pattern is similar to that for a single point source, and it is still possible to tell that there are two light sources rather than one. If they are closer together, as in [link](c), we cannot distinguish them, thus limiting the detail or **resolution** we can obtain. This limit is an inescapable consequence of the wave nature of light.

Diffraction limits the resolution in many situations. The acuity of our vision is limited because light passes through the pupil, which is the circular aperture of the eye. Be aware that the diffraction-like spreading of light is due to the limited diameter of a light beam, not the interaction with an aperture. Thus, light passing through a lens with a diameter *D* shows this effect and spreads, blurring the image, just as light passing through an aperture of diameter *D* does. Thus, diffraction limits the resolution of any system having a lens or mirror. Telescopes are also limited by diffraction, because of the finite diameter *D* of the primary mirror.

Just what is the limit? To answer that question, consider the diffraction pattern for a circular aperture, which has a central maximum that is wider and brighter than the maxima surrounding it (similar to a slit) ([link](a)). It can be shown that, for a circular aperture of diameter *D*, the first minimum in the diffraction pattern occurs at $\theta = 1.22\lambda/D$ (providing the aperture is large compared with the wavelength of light, which is the case for most optical instruments). The accepted criterion for determining the **diffraction limit** to resolution based on this angle is known as the **Rayleigh criterion**, which was developed by Lord Rayleigh in the nineteenth century.

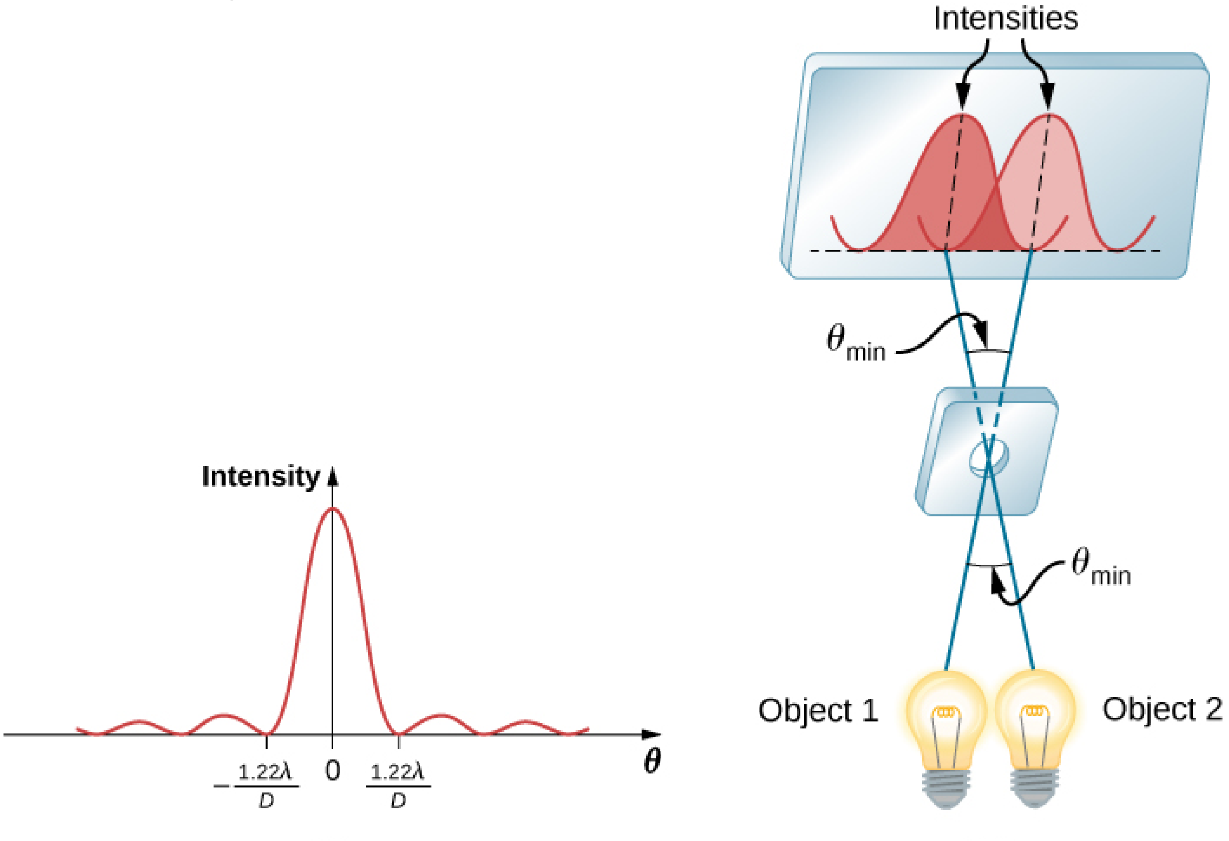
Note:
Rayleigh Criterion
The diffraction limit to resolution states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other ([link](b)).

The first minimum is at an angle of $\theta = 1.22\lambda/D$, so that two point objects are just resolvable if they are separated by the angle

Note:
Equation:

$$\theta = 1.22\frac{\lambda}{D}$$

where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc., with which the two objects are observed. In this expression, θ has units of radians. This angle is also commonly known as the diffraction limit.


Intensity
θ
$-\frac{1.22\lambda}{D}$
0
$\frac{1.22\lambda}{D}$
Intensities
θ_{min}
θ_{min}
Object 1
Object 2

(a)

(b)

(a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Even the small wavelength of light prohibits exact precision. When extremely small wavelength probes are used, as with an electron microscope, the system is disturbed, still limiting our knowledge. Heisenberg’s uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in the chapter on quantum mechanics.

Example:

Calculating Diffraction Limits of the Hubble Space Telescope

The primary mirror of the orbiting Hubble Space Telescope has a diameter of 2.40 m. Being in orbit, this telescope avoids the degrading effects of atmospheric distortion on its resolution. (a) What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm. (b) If these two stars are at a distance of 2 million light-years, which is the distance of the Andromeda Galaxy, how close together can they be and still be resolved? (A light-year, or ly, is the distance light travels in 1 year.)

Strategy

The Rayleigh criterion stated in [link], $\theta = 1.22\lambda / D$, gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is known, we can calculate the distance between the stars, since we are given how far away they are.

Solution

a. The Rayleigh criterion for the minimum resolvable angle is
   Equation:

$$\theta = 1.22\frac{\lambda}{D}.$$

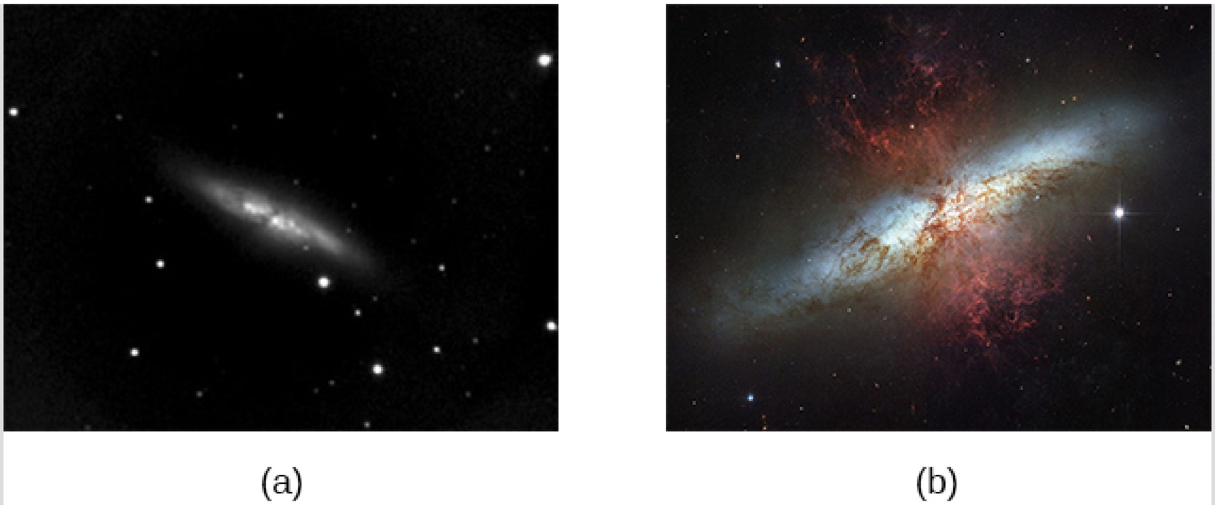
Entering known values gives
Equation:

$$\theta = 1.22\frac{550\ \times\ 10^{-9}\,\text{m}}{2.40\,\text{m}} = 2.80\ \times\ 10^{-7}\,\text{rad}.$$

b. The distance *s* between two objects a distance *r* away and separated by an angle θ is $s = r\theta$.
Substituting known values gives
Equation:

$$s = \left(2.0\ \times\ 10^{6}\,\text{ly}\right)\left(2.80\ \times\ 10^{-7}\,\text{rad}\right) = 0.56\,\text{ly}.$$

Significance
The angle found in part (a) is extraordinarily small (less than 1/50,000 of a degree), because the primary mirror is so large compared with the wavelength of light. As noticed, diffraction effects are most noticeable when light interacts with objects having sizes on the order of the wavelength of light. However, the effect is still there, and there is a diffraction limit to what is observable. The actual resolution of the Hubble Telescope is not quite as good as that found here. As with all instruments, there are other effects, such as nonuniformities in mirrors or aberrations in lenses that further limit resolution. However, [link] gives an indication of the extent of the detail observable with the Hubble because of its size and quality, and especially because it is above Earth's atmosphere.


(a) (b)

These two photographs of the M82 Galaxy give an idea of the observable detail using (a) a ground-based telescope and (b) the Hubble Space Telescope. (credit a: modification of work by “Ricnun”/Wikimedia Commons; credit b: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA))

The answer in part (b) indicates that two stars separated by about half a light-year can be resolved. The average distance between stars in a galaxy is on the order of five light-years in the outer parts and about one light-year near the galactic center. Therefore, the Hubble can resolve most of the individual stars in Andromeda Galaxy, even though it lies at such a huge distance that its light takes 2 million years to reach us. [link] shows another mirror used to observe radio waves from outer space.

A 305-m-diameter paraboloid at Arecibo in Puerto Rico is lined with reflective material, making it into a radio telescope. It is the largest curved focusing dish in the world. Although *D* for Arecibo is much larger than for the Hubble Telescope, it detects radiation of a much longer wavelength and its diffraction limit is significantly poorer than Hubble’s. The Arecibo telescope is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Jeff Hitchcock)

Note:
Exercise:

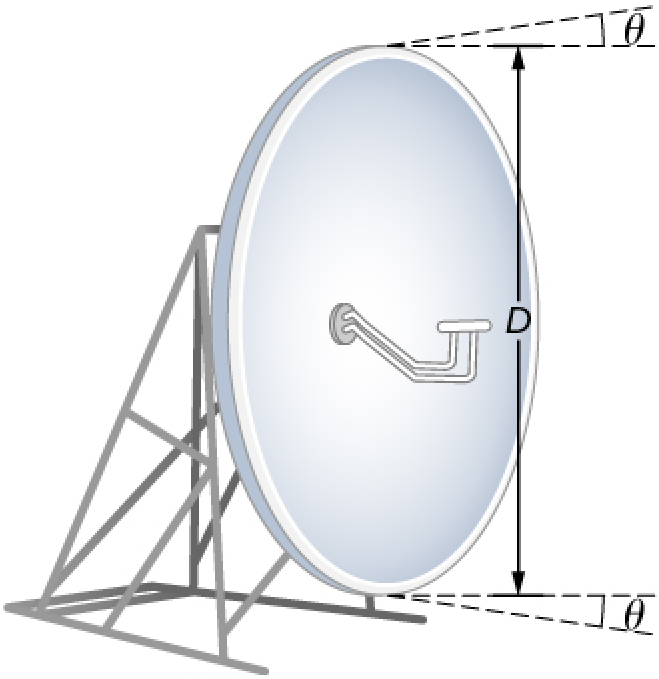
Problem:

Check Your Understanding What is the angular resolution of the Arecibo telescope shown in [link] when operated at 21-cm wavelength? How does it compare to the resolution of the Hubble Telescope?

Solution:

$8.4\ \times\ 10^{-4}$ rad, 3000 times broader than the Hubble Telescope

Diffraction is not only a problem for optical instruments but also for the electromagnetic radiation itself. Any beam of light having a finite diameter *D* and a wavelength λ exhibits diffraction spreading. The beam spreads out with an angle θ given by [link], $\theta = 1.22\lambda/D$. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to $\theta = 0^\circ$ as possible) instead spreads out at an angle $\theta = 1.22\lambda/D$, where D is the diameter of the beam and λ is its wavelength. This spreading is impossible to observe for a flashlight because its beam is not very parallel to start with. However, for long-distance transmission of laser beams or microwave signals, diffraction spreading can be significant ([link]). To avoid this, we can increase *D*. This is done for laser light sent to the moon to measure its distance from Earth. The laser beam is expanded through a telescope to make *D* much larger and θ smaller.



The beam produced by this microwave transmission antenna spreads out at a minimum angle $\theta = 1.22\lambda/D$ due to diffraction. It is impossible to produce a near-parallel beam because the beam has a limited diameter.

In most biology laboratories, resolution is an issue when the use of the microscope is introduced. The smaller the distance *x* by which two objects can be separated and still be seen as distinct, the greater the resolution. The resolving power of a lens is defined as that distance *x*. An expression for resolving power is obtained from the Rayleigh criterion. [link](a) shows two point objects separated by a distance *x*. According to the Rayleigh criterion, resolution is possible when the minimum angular separation is

Equation:

$$\theta = 1.22\frac{\lambda}{D} = \frac{x}{d},$$

where *d* is the distance between the specimen and the objective lens, and we have used the small angle approximation (i.e., we have assumed that *x* is much smaller than *d*), so that $\tan\theta \approx \sin\theta \approx \theta$. Therefore, the resolving power is

Equation:

$$x = 1.22\frac{\lambda d}{D}.$$

Another way to look at this is by the concept of numerical aperture (*NA*), which is a measure of the maximum acceptance angle at which a lens will take light and still contain it within the lens. [link](b) shows a lens and an object at point *P*. The *NA* here is a measure of the ability of the lens to gather light and resolve fine detail. The angle subtended by the lens at its focus is defined to be $\theta = 2\alpha$. From the figure and again using the small angle approximation, we can write

Equation:

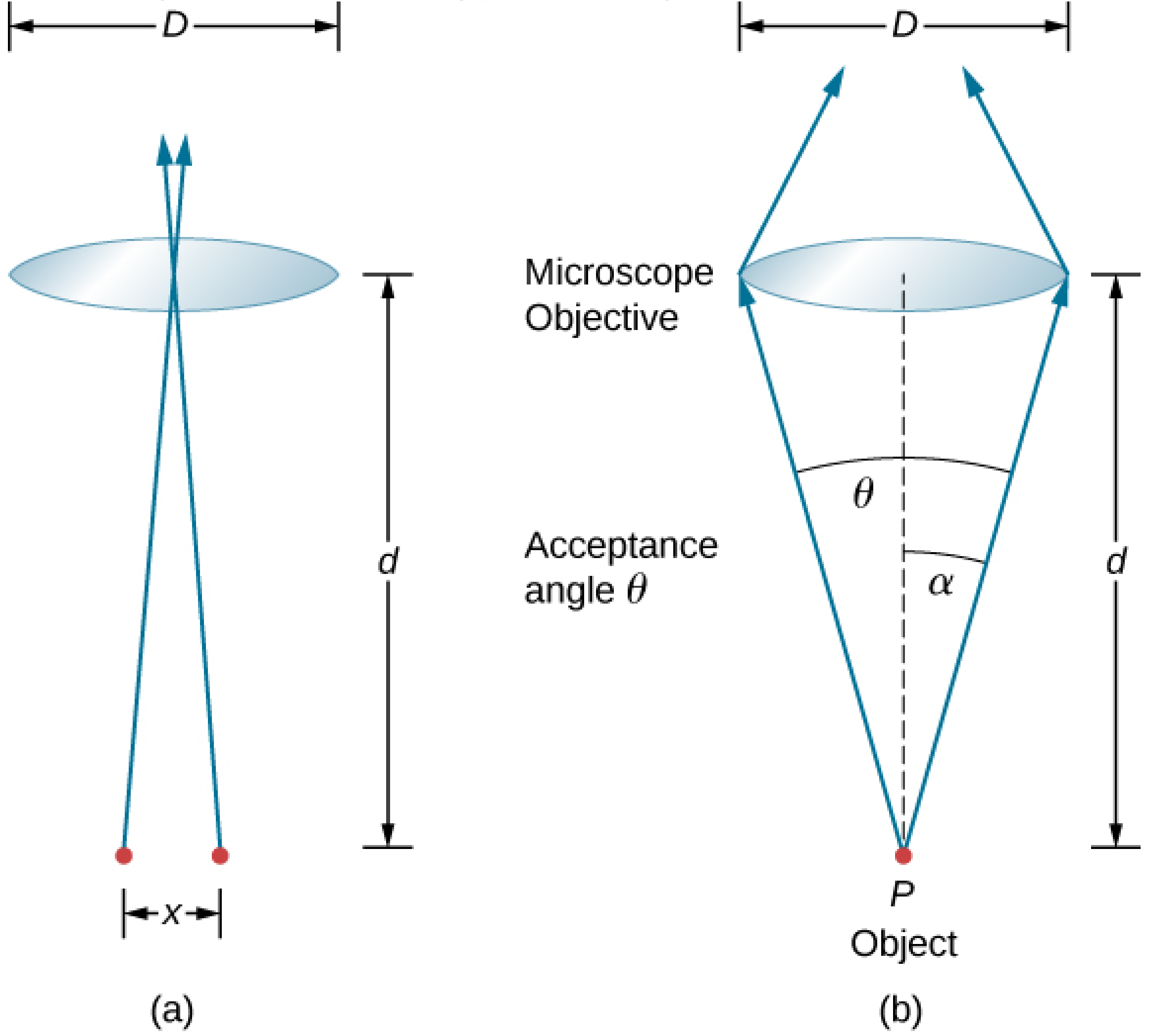
$$\sin\alpha = \frac{D/2}{d} = \frac{D}{2d}.$$

The *NA* for a lens is $NA = n\sin\alpha$, where *n* is the index of refraction of the medium between the objective lens and the object at point *P*. From this definition for *NA*, we can see that

Equation:

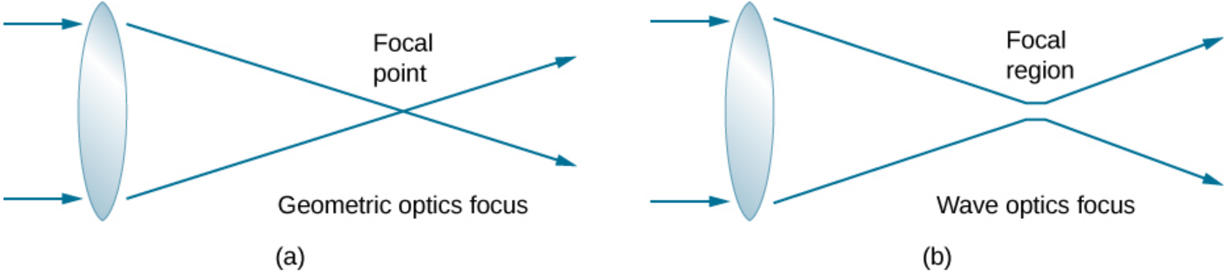
$$x = 1.22\frac{\lambda d}{D} = 1.22\frac{\lambda}{2\sin\alpha} = 0.61\frac{\lambda n}{NA}.$$

In a microscope, *NA* is important because it relates to the resolving power of a lens. A lens with a large *NA* is able to resolve finer details. Lenses with larger *NA* are also able to collect more light and so give a brighter image. Another way to describe this situation is that the larger the *NA*, the larger the cone of light that can be brought into the lens, so more of the diffraction modes are collected. Thus the microscope has more information to form a clear image, and its resolving power is higher.



(a) Two points separated by a distance x and positioned a distance d away from the objective. (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P (credit a: modification of work by “Infopro”/Wikimedia Commons).

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Imagine focusing when only considering geometric optics, as in [link](a). The focal point is regarded as an infinitely small point with a huge intensity and the capacity to incinerate most samples, irrespective of the *NA* of the objective lens—an unphysical oversimplification. For wave optics, due to diffraction, we take into account the phenomenon in which the focal point spreads to become a focal spot ([link](b)) with the size of the spot decreasing with increasing *NA*. Consequently, the intensity in the focal spot increases with increasing *NA*. The higher the *NA*, the greater the chances of photodegrading the specimen. However, the spot never becomes a true point.



(a) In geometric optics, the focus is modelled as a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

In a different type of microscope, molecules within a specimen are made to emit light through a mechanism called fluorescence. By controlling the molecules emitting light, it has become possible to construct images with resolution much finer than the Rayleigh criterion, thus circumventing the diffraction limit. The development of super-resolved fluorescence microscopy led to the 2014 Nobel Prize in Chemistry.

Note:
In this Optical Resolution Model, two diffraction patterns for light through two circular apertures are shown side by side in this simulation by Fu-Kwun Hwang. Watch the patterns merge as you decrease the aperture diameters.

Summary

- Diffraction limits resolution.
- The Rayleigh criterion states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other.

Conceptual Questions

Exercise:

Problem:

Is higher resolution obtained in a microscope with red or blue light? Explain your answer.

Solution:

blue; The shorter wavelength of blue light results in a smaller angle for diffraction limit.

Exercise:

Problem:

The resolving power of refracting telescope increases with the size of its objective lens. What other advantage is gained with a larger lens?

Exercise:

Problem:

The distance between atoms in a molecule is about $10^{-8}\ \text{cm}$. Can visible light be used to “see” molecules?

Solution:

No, these distances are three orders of magnitude smaller than the wavelength of visible light, so visible light makes a poor probe for atoms.

Exercise:

Problem:

A beam of light always spreads out. Why can a beam not be created with parallel rays to prevent spreading? Why can lenses, mirrors, or apertures not be used to correct the spreading?

Problems

Exercise:

Problem:

The 305-m-diameter Arecibo radio telescope pictured in [link] detects radio waves with a 4.00-cm average wavelength. (a) What is the angle between two just-resolvable point sources for this telescope? (b) How close together could these point sources be at the 2 million light-year distance of the Andromeda Galaxy?

Exercise:

Problem:

Assuming the angular resolution found for the Hubble Telescope in [link], what is the smallest detail that could be observed on the moon?

Solution:

107 m

Exercise:

Problem:

Diffraction spreading for a flashlight is insignificant compared with other limitations in its optics, such as spherical aberrations in its mirror. To show this, calculate the minimum angular spreading of a flashlight beam that is originally 5.00 cm in diameter with an average wavelength of 600 nm.

Exercise:

Problem:

(a) What is the minimum angular spread of a 633-nm wavelength He-Ne laser beam that is originally 1.00 mm in diameter? (b) If this laser is aimed at a mountain cliff 15.0 km away, how big will the illuminated spot be? (c) How big a spot would be illuminated on the moon, neglecting atmospheric effects? (This might be done to hit a corner reflector to measure the round-trip time and, hence, distance.)

Solution:

a. $7.72 \times 10^{-4}\,\text{rad}$; b. 23.2 m; c. 590 km

Exercise:

Problem:

A telescope can be used to enlarge the diameter of a laser beam and limit diffraction spreading. The laser beam is sent through the telescope in opposite the normal direction and can then be projected onto a satellite or the moon. (a) If this is done with the Mount Wilson telescope, producing a 2.54-m-diameter beam of 633-nm light, what is the minimum angular spread of the beam? (b) Neglecting atmospheric effects, what is the size of the spot this beam would make on the moon, assuming a lunar distance of $3.84 \times 10^{8}\,\text{m}$?

Exercise:

Problem:

The limit to the eye’s acuity is actually related to diffraction by the pupil. (a) What is the angle between two just-resolvable points of light for a 3.00-mm-diameter pupil, assuming an average wavelength of 550 nm? (b) Take your result to be the practical limit for the eye. What is the greatest possible distance a car can be from you if you can resolve its two headlights, given they are 1.30 m apart? (c) What is the distance between two just-resolvable points held at an arm’s length (0.800 m) from your eye? (d) How does your answer to (c) compare to details you normally observe in everyday circumstances?

Solution:

a. 2.24×10^{-4} rad; b. 5.81 km; c. 0.179 mm; d. can resolve details 0.2 mm apart at arm’s length

Exercise:

Problem:

What is the minimum diameter mirror on a telescope that would allow you to see details as small as 5.00 km on the moon some 384,000 km away? Assume an average wavelength of 550 nm for the light received.

Exercise:

Problem:

Find the radius of a star’s image on the retina of an eye if its pupil is open to 0.65 cm and the distance from the pupil to the retina is 2.8 cm. Assume $\lambda = 550\,\text{nm}$.

Solution:

$2.9\ \mu\text{m}$

Exercise:

Problem:

(a) The dwarf planet Pluto and its moon, Charon, are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Palomar Mountain telescope be able to resolve these bodies when they are $4.50 \times 10^9\,\text{km}$ from Earth? Assume an average wavelength of 550 nm. (b) In actuality, it is just barely possible to discern that Pluto and Charon are separate bodies using a ground-based telescope. What are the reasons for this?

Exercise:

Problem:

A spy satellite orbits Earth at a height of 180 km. What is the minimum diameter of the objective lens in a telescope that must be used to resolve columns of troops marching 2.0 m apart? Assume $\lambda = 550\,\text{nm}$.

Solution:

6.0 cm

Exercise:

Problem:

What is the minimum angular separation of two stars that are just-resolvable by the 8.1-m Gemini South telescope, if atmospheric effects do not limit resolution? Use 550 nm for the wavelength of the light from the stars.

Exercise:

Problem:

The headlights of a car are 1.3 m apart. What is the maximum distance at which the eye can resolve these two headlights? Take the pupil diameter to be 0.40 cm.

Solution:

7.71 km

Exercise:

Problem:

When dots are placed on a page from a laser printer, they must be close enough so that you do not see the individual dots of ink. To do this, the separation of the dots must be less than Raleigh’s criterion. Take the pupil of the eye to be 3.0 mm and the distance from the paper to the eye of 35 cm; find the minimum separation of two dots such that they cannot be resolved. How many dots per inch (dpi) does this correspond to?

Exercise:

Problem:

Suppose you are looking down at a highway from a jetliner flying at an altitude of 6.0 km. How far apart must two cars be if you are able to distinguish them? Assume that $\lambda = 550 \text{ nm}$ and that the diameter of your pupils is 4.0 mm.

Solution:

1.0 m

Exercise:

Problem:

Can an astronaut orbiting Earth in a satellite at a distance of 180 km from the surface distinguish two skyscrapers that are 20 m apart? Assume that the pupils of the astronaut’s eyes have a diameter of 5.0 mm and that most of the light is centered around 500 nm.

Exercise:

Problem:

The characters of a stadium scoreboard are formed with closely spaced lightbulbs that radiate primarily yellow light. (Use $\lambda = 600\ \mathrm{nm}$.) How closely must the bulbs be spaced so that an observer 80 m away sees a display of continuous lines rather than the individual bulbs? Assume that the pupil of the observer's eye has a diameter of 5.0 mm.

Solution:

1.2 cm or closer

Exercise:

Problem:

If a microscope can accept light from objects at angles as large as $\alpha = 70^\circ$, what is the smallest structure that can be resolved when illuminated with light of wavelength 500 nm and (a) the specimen is in air? (b) When the specimen is immersed in oil, with index of refraction of 1.52?

Exercise:

Problem:

A camera uses a lens with aperture 2.0 cm. What is the angular resolution of a photograph taken at 700 nm wavelength? Can it resolve the millimeter markings of a ruler placed 35 m away?

Solution:

no

Glossary

diffraction limit
: fundamental limit to resolution due to diffraction

Rayleigh criterion

two images are just-resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other

resolution

ability, or limit thereof, to distinguish small details in images

X-Ray Diffraction

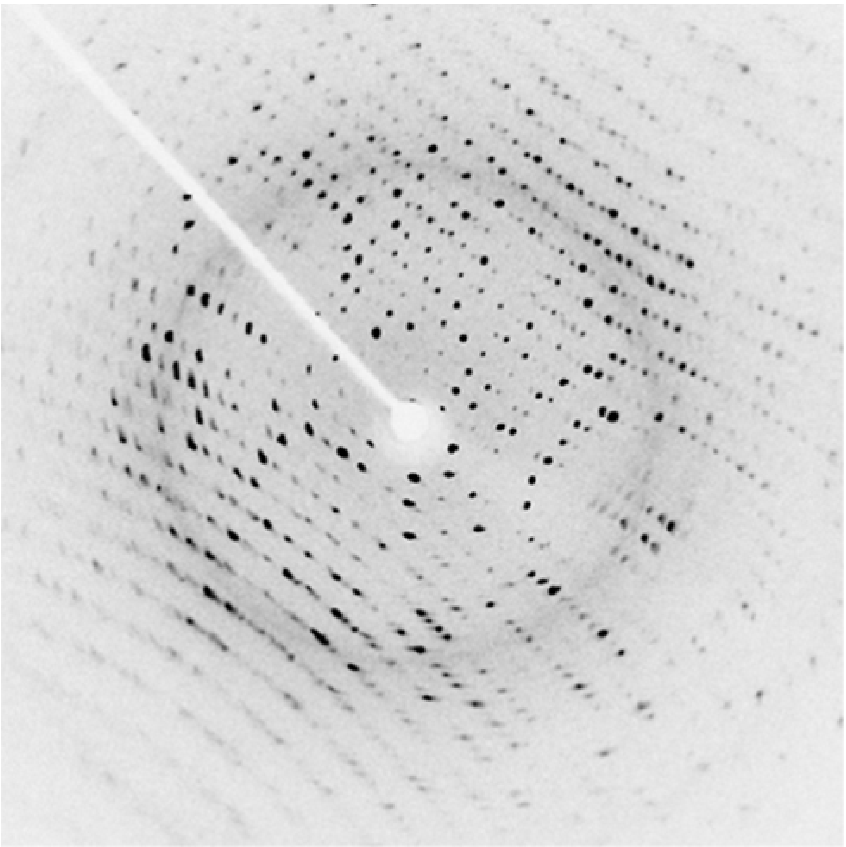
By the end of this section, you will be able to:

- Describe interference and diffraction effects exhibited by X-rays in interaction with atomic-scale structures

Since X-ray photons are very energetic, they have relatively short wavelengths, on the order of 10^{-8} m to 10^{-12} m. Thus, typical X-ray photons act like rays when they encounter macroscopic objects, like teeth, and produce sharp shadows. However, since atoms are on the order of 0.1 nm in size, X-rays can be used to detect the location, shape, and size of atoms and molecules. The process is called **X-ray diffraction**, and it involves the interference of X-rays to produce patterns that can be analyzed for information about the structures that scattered the X-rays.

Perhaps the most famous example of X-ray diffraction is the discovery of the double-helical structure of DNA in 1953 by an international team of scientists working at England’s Cavendish Laboratory—American James Watson, Englishman Francis Crick, and New Zealand-born Maurice Wilkins. Using X-ray diffraction data produced by Rosalind Franklin, they were the first to model the double-helix structure of DNA that is so crucial to life. For this work, Watson, Crick, and Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine. (There is some debate and controversy over the issue that Rosalind Franklin was not included in the prize, although she died in 1958, before the prize was awarded.)

[link] shows a diffraction pattern produced by the scattering of X-rays from a crystal. This process is known as X-ray crystallography because of the information it can yield about crystal structure, and it was the type of data Rosalind Franklin supplied to Watson and Crick for DNA. Not only do X-rays confirm the size and shape of atoms, they give information about the atomic arrangements in materials. For example, more recent research in high-temperature superconductors involves complex materials whose lattice arrangements are crucial to obtaining a superconducting material. These can be studied using X-ray crystallography.



X-ray diffraction from the crystal of a protein (hen egg lysozyme) produced this interference pattern. Analysis of the pattern yields information about the structure of the protein. (credit: “Del45”/Wikimedia Commons)

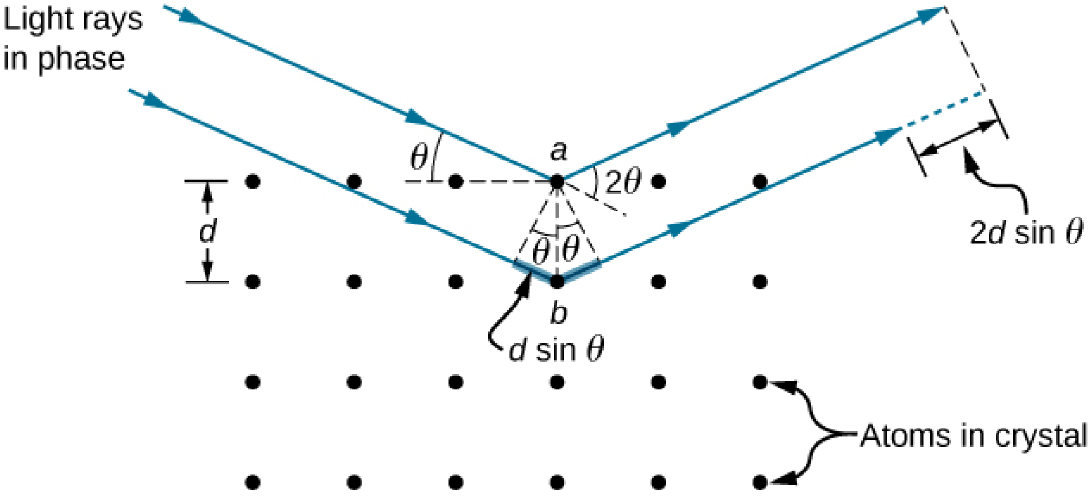
Historically, the scattering of X-rays from crystals was used to prove that X-rays are energetic electromagnetic (EM) waves. This was suspected from the time of the discovery of X-rays in 1895, but it was not until 1912 that the German Max von Laue (1879–1960) convinced two of his colleagues to scatter X-rays from crystals. If a diffraction pattern is obtained, he reasoned, then the X-rays must be waves, and their wavelength could be determined. (The spacing of atoms in various crystals was reasonably well known at the time, based on good values for Avogadro’s number.) The experiments were convincing, and the 1914 Nobel Prize in Physics was given to von Laue for his suggestion leading to the proof that X-rays are EM waves. In 1915, the unique father-and-son team of Sir William Henry Bragg and his son Sir William Lawrence Bragg were awarded a joint Nobel Prize for inventing the X-ray spectrometer and the then-new science of X-ray analysis.

In ways reminiscent of thin-film interference, we consider two plane waves at X-ray wavelengths, each one reflecting off a different plane of atoms within a crystal’s lattice, as shown in [link]. From the geometry, the difference in path lengths is $2d\sin\theta$. Constructive interference results when this distance is an integer multiple of the wavelength. This condition is captured by the *Bragg equation*,

Note:
Equation:

$$m\lambda = 2d\sin\theta,\ m = 1,\ 2,\ 3\ldots$$

where *m* is a positive integer and *d* is the spacing between the planes. Following the Law of Reflection, both the incident and reflected waves are described by the same angle, θ, but unlike the general practice in geometric optics, θ is measured with respect to the surface itself, rather than the normal.



X-ray diffraction with a crystal. Two incident waves reflect off two planes of a crystal. The difference in path lengths is indicated by the dashed line.

Example:
X-Ray Diffraction with Salt Crystals

Common table salt is composed mainly of NaCl crystals. In a NaCl crystal, there is a family of planes 0.252 nm apart. If the first-order maximum is observed at an incidence angle of 18.1°, what is the wavelength of the X-ray scattering from this crystal?

Strategy

Use the Bragg equation, [link], $m\lambda = 2d\sin\theta$, to solve for θ.

Solution

For first-order, $m = 1$, and the plane spacing *d* is known. Solving the Bragg equation for wavelength yields

Equation:

$$\lambda = \frac{2d\sin\theta}{m} = \frac{2\left(0.252\ \times\ 10^{-9}\,\text{m}\right)\sin\ (18.1^\circ)}{1} = 1.57\ \times\ 10^{-10}\,\text{m},\ \text{or}\ 0.157\,\text{nm}.$$

Significance

The determined wavelength fits within the X-ray region of the electromagnetic spectrum. Once again, the wave nature of light makes itself prominent when the wavelength $(\lambda = 0.157\,\text{nm})$ is comparable to the size of the physical structures $(d = 0.252\,\text{nm})$ it interacts with.

Note:

Exercise:

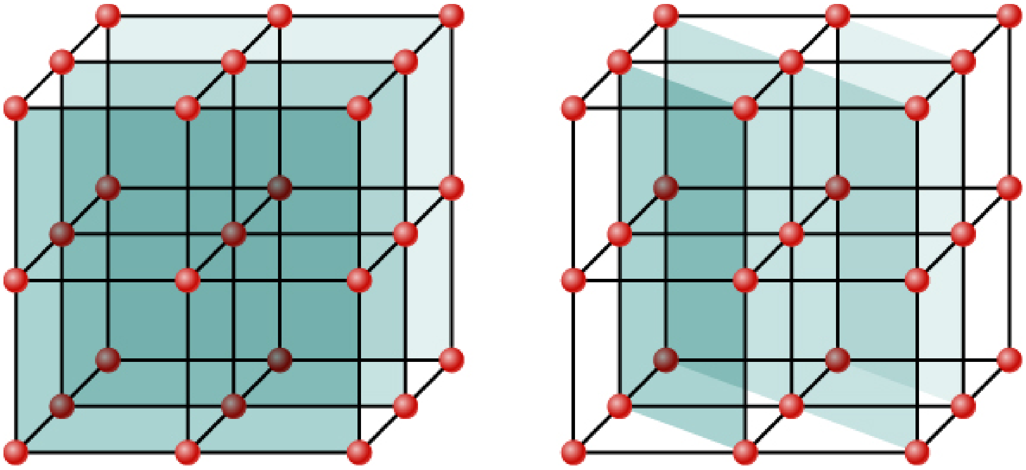
Problem:

Check Your Understanding For the experiment described in [link], what are the two other angles where interference maxima may be observed? What limits the number of maxima?

Solution:

38.4° and 68.8°; Between $\theta = 0^\circ \rightarrow 90^\circ$, orders 1, 2, and 3, are all that exist.

Although [link] depicts a crystal as a two-dimensional array of scattering centers for simplicity, real crystals are structures in three dimensions. Scattering can occur simultaneously from different families of planes at different orientations and spacing patterns known as called **Bragg planes**, as shown in [link]. The resulting interference pattern can be quite complex.



Because of the regularity that makes a crystal structure, one crystal can have many families of planes within its geometry, each one giving rise to X-ray diffraction.

Summary

- X-rays are relatively short-wavelength EM radiation and can exhibit wave characteristics such as interference when interacting with correspondingly small objects.

Conceptual Questions

Exercise:

Problem: Crystal lattices can be examined with X-rays but not UV. Why?

Solution:

UV wavelengths are much larger than lattice spacings in crystals such that there is no diffraction. The Bragg equation implies a value for $sin\theta$ greater than unity, which has no solution.

Problems

Exercise:

Problem:

X-rays of wavelength 0.103 nm reflects off a crystal and a second-order maximum is recorded at a Bragg angle of 25.5°. What is the spacing between the scattering planes in this crystal?

Exercise:

Problem:

A first-order Bragg reflection maximum is observed when a monochromatic X-ray falls on a crystal at a 32.3° angle to a reflecting plane. What is the wavelength of this X-ray?

Solution:

0.120 nm

Exercise:

Problem:

An X-ray scattering experiment is performed on a crystal whose atoms form planes separated by 0.440 nm. Using an X-ray source of wavelength 0.548 nm, what is the angle (with respect to the planes in question) at which the experimenter needs to illuminate the crystal in order to observe a first-order maximum?

Exercise:

Problem:

The structure of the NaCl crystal forms reflecting planes 0.541 nm apart. What is the smallest angle, measured from these planes, at which X-ray diffraction can be observed, if X-rays of wavelength 0.085 nm are used?

Solution:

4.51°

Exercise:

Problem:

On a certain crystal, a first-order X-ray diffraction maximum is observed at an angle of 27.1° relative to its surface, using an X-ray source of unknown wavelength. Additionally, when illuminated with a different, this time of known wavelength 0.137 nm, a second-order maximum is detected at 37.3°. Determine (a) the spacing between the reflecting planes, and (b) the unknown wavelength.

Exercise:

Problem:

Calcite crystals contain scattering planes separated by 0.30 nm. What is the angular separation between first and second-order diffraction maxima when X-rays of 0.130 nm wavelength are used?

Solution:

13.2°

Exercise:

Problem:

The first-order Bragg angle for a certain crystal is 12.1°. What is the second-order angle?

Glossary

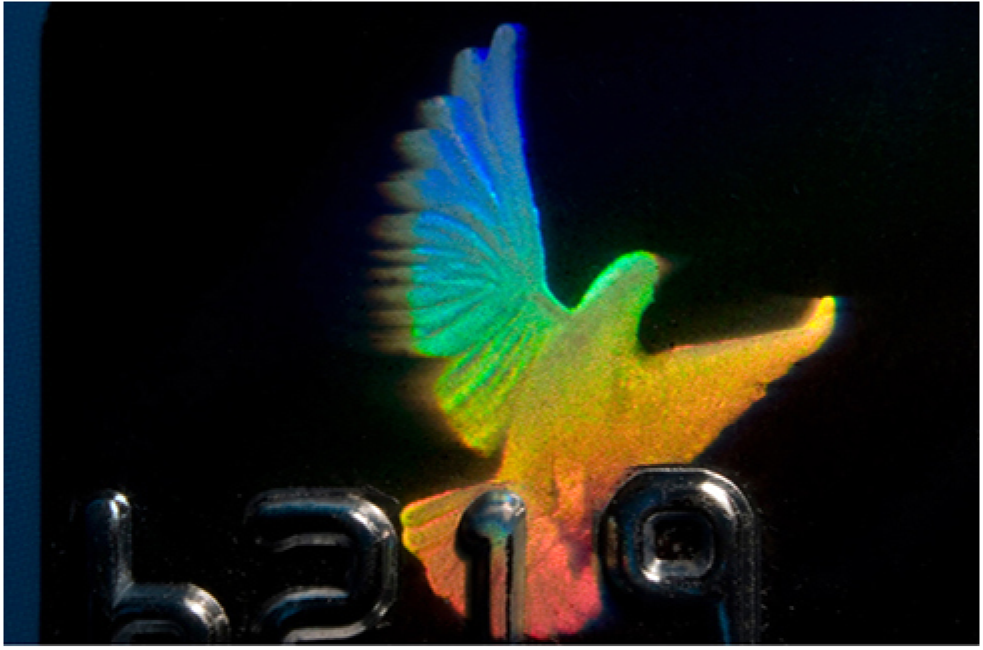
Bragg planes	families of planes within crystals that can give rise to X-ray diffraction
X-ray diffraction	technique that provides the detailed information about crystallographic structure of natural and manufactured materials

Holography

By the end of this section, you will be able to:

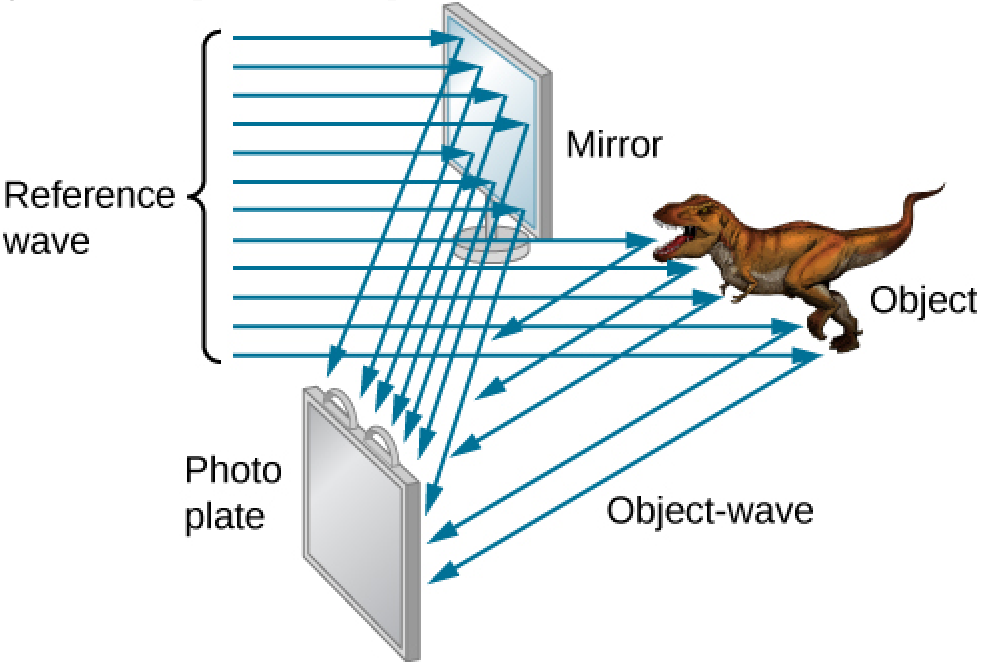
- Describe how a three-dimensional image is recorded as a hologram
- Describe how a three-dimensional image is formed from a hologram

A **hologram**, such as the one in [link], is a true three-dimensional image recorded on film by lasers. Holograms are used for amusement; decoration on novelty items and magazine covers; security on credit cards and driver's licenses (a laser and other equipment are needed to reproduce them); and for serious three-dimensional information storage. You can see that a hologram is a true three-dimensional image because objects change relative position in the image when viewed from different angles.



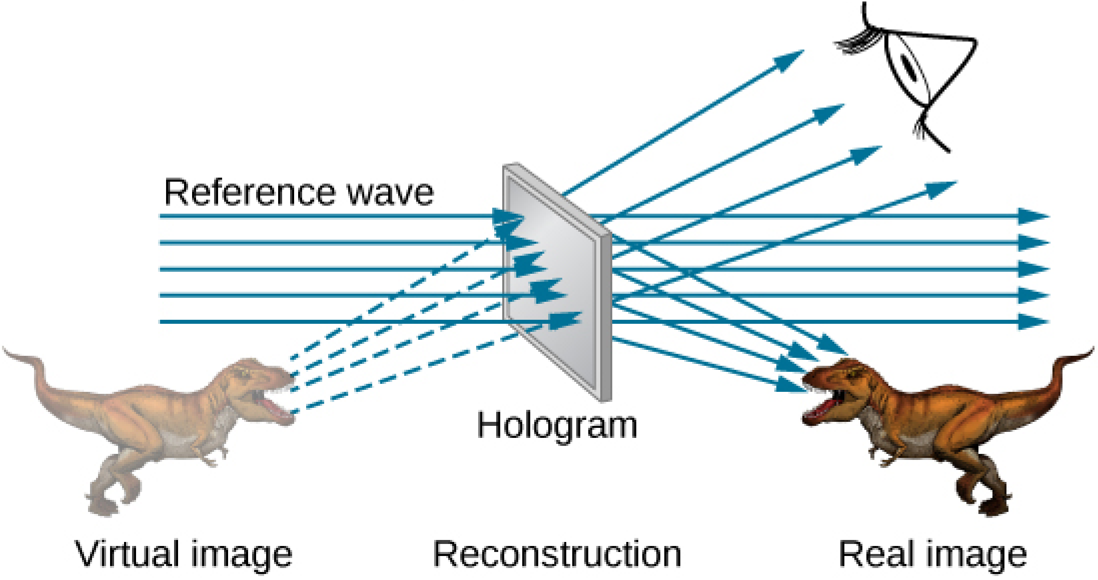
Credit cards commonly have holograms for logos, making them difficult to reproduce. (credit: Dominic Alves)

The name hologram means "entire picture" (from the Greek *holo*, as in holistic) because the image is three-dimensional. **Holography** is the process of producing holograms and, although they are recorded on photographic film, the process is quite different from normal photography. Holography uses light interference or wave optics, whereas normal photography uses geometric optics. [link] shows one method of producing a hologram. Coherent light from a laser is split by a mirror, with part of the light illuminating the object. The remainder, called the reference beam, shines directly on a piece of film. Light scattered from the object interferes with the reference beam, producing constructive and destructive interference. As a result, the exposed film looks foggy, but close examination reveals a complicated interference pattern stored on it. Where the interference was constructive, the film (a negative actually) is darkened. Holography is sometimes called lens-less photography, because it uses the wave characteristics of light, as contrasted to normal photography, which uses geometric optics and requires lenses.



Production of a hologram. Single-wavelength coherent light from a laser produces a well-defined interference pattern on a piece of film. The laser beam is split by a partially silvered mirror, with part of the light illuminating the object and the remainder shining directly on the film. (credit: modification of work by Mariana Ruiz Villarreal)

Light falling on a hologram can form a three-dimensional image of the original object. The process is complicated in detail, but the basics can be understood, as shown in [link], in which a laser of the same type that exposed the film is now used to illuminate it. The myriad tiny exposed regions of the film are dark and block the light, whereas less exposed regions allow light to pass. The film thus acts much like a collection of diffraction gratings with various spacing patterns. Light passing through the hologram is diffracted in various directions, producing both real and virtual images of the object used to expose the film. The interference pattern is the same as that produced by the object. Moving your eye to various places in the interference pattern gives you different perspectives, just as looking directly at the object would. The image thus looks like the object and is three dimensional like the object.



A transmission hologram is one that produces real and virtual images when a laser of the same type as that which exposed the hologram is passed through it. Diffraction from various parts of the film produces the same interference pattern that was produced by the object that was used to expose it. (credit: modification of work by Mariana Ruiz Villarreal)

The hologram illustrated in [link] is a transmission hologram. Holograms that are viewed with reflected light, such as the white light holograms on credit cards, are reflection holograms and are more common. White light holograms often appear a little blurry with rainbow edges, because the diffraction patterns of various colors of light are at slightly different locations due to their different wavelengths. Further uses of holography include all types of three-dimensional information storage, such as of statues in museums, engineering studies of structures, and images of human organs.

Invented in the late 1940s by Dennis Gabor (1900–1970), who won the 1971 Nobel Prize in Physics for his work, holography became far more practical with the development of the laser. Since lasers produce coherent single-wavelength light, their interference patterns are more pronounced. The precision is so great that it is even possible to record numerous holograms on a single piece of film by just changing the angle of the film for each successive image. This is how the holograms that move as you walk by them are produced—a kind of lens-less movie.

In a similar way, in the medical field, holograms have allowed complete three-dimensional holographic displays of objects from a stack of images. Storing these images for future use is relatively easy. With the use of an endoscope, high-resolution, three-dimensional holographic images of internal organs and tissues can be made.

Summary

- Holography is a technique based on wave interference to record and form three-dimensional images.
- Lasers offer a practical way to produce sharp holographic images because of their monochromatic and coherent light for pronounced interference patterns.

Key Equations

Destructive interference for a single slit	$D\sin\theta = m\lambda \text{ for } m = \pm 1,\ \pm 2,\ \pm 3, \ldots$
Half phase angle	$\beta = \frac{\phi}{2} = \frac{\pi D\sin\theta}{\lambda}$
Field amplitude in the diffraction pattern	$E = N\Delta E_0 \frac{\sin\beta}{\beta}$
Intensity in the diffraction pattern	$I = I_0 \left(\frac{\sin\beta}{\beta}\right)^2$
Rayleigh criterion for circular apertures	$\theta = 1.22\frac{\lambda}{D}$
Bragg equation	$m\lambda = 2d\sin\theta,\ m = 1,\ 2,\ 3\ldots$

Conceptual Questions

Exercise:

Problem:

How can you tell that a hologram is a true three-dimensional image and that those in three-dimensional movies are not?

Exercise:

Problem:

If a hologram is recorded using monochromatic light at one wavelength but its image is viewed at another wavelength, say 10% shorter, what will you see? What if it is viewed using light of exactly half the original wavelength?

Solution:

Image will appear at slightly different location and/or size when viewed using 10% shorter wavelength but at exactly half the wavelength, a higher-order interference reconstructs the original image, different color.

Exercise:

Problem:

What image will one see if a hologram is recorded using monochromatic light but its image is viewed in white light? Explain.

Additional Problems

Exercise:

Problem:

White light falls on two narrow slits separated by 0.40 mm. The interference pattern is observed on a screen 3.0 m away. (a) What is the separation between the first maxima for red light $(\lambda = 700\ \text{nm})$ and violet light $(\lambda = 400\ \text{nm})$? (b) At what point nearest the central maximum will a maximum for yellow light $(\lambda = 600\ \text{nm})$ coincide with a maximum for violet light? Identify the order for each maximum.

Solution:

a. 2.2 mm; b. 0.172°, second-order yellow and third-order violet coincide

Exercise:

Problem:

Microwaves of wavelength 10.0 mm fall normally on a metal plate that contains a slit 25 mm wide. (a) Where are the first minima of the diffraction pattern? (b) Would there be minima if the wavelength were 30.0 mm?

Exercise:

Problem:

Quasars, or *quasi-stellar radio sources*, are astronomical objects discovered in 1960. They are distant but strong emitters of radio waves with angular size so small, they were originally unresolved, the same as stars. The quasar 3C405 is actually two discrete radio sources that subtend an angle of 82 arcsec. If this object is studied using radio emissions at a frequency of 410 MHz, what is the minimum diameter of a radio telescope that can resolve the two sources?

Solution:

2.2 km

Exercise:

Problem:

Two slits each of width 1800 nm and separated by the center-to-center distance of 1200 nm are illuminated by plane waves from a krypton ion laser-emitting at wavelength 461.9 nm. Find the number of interference peaks in the central diffraction peak.

Exercise:

Problem:

A microwave of an unknown wavelength is incident on a single slit of width 6 cm. The angular width of the central peak is found to be 25°. Find the wavelength.

Solution:

1.3 cm

Exercise:

Problem:

Red light (wavelength 632.8 nm in air) from a Helium-Neon laser is incident on a single slit of width 0.05 mm. The entire apparatus is immersed in water of refractive index 1.333. Determine the angular width of the central peak.

Exercise:

Problem:

A light ray of wavelength 461.9 nm emerges from a 2-mm circular aperture of a krypton ion laser. Due to diffraction, the beam expands as it moves out. How large is the central bright spot at (a) 1 m, (b) 1 km, (c) 1000 km, and (d) at the surface of the moon at a distance of 400,000 km from Earth.

Solution:

a. 0.28 mm; b. 0.28 m; c. 280 m; d. 113 km

Exercise:

Problem:

How far apart must two objects be on the moon to be distinguishable by eye if only the diffraction effects of the eye's pupil limit the resolution? Assume 550 nm for the wavelength of light, the pupil diameter 5.0 mm, and 400,000 km for the distance to the moon.

Exercise:

Problem:

How far apart must two objects be on the moon to be resolvable by the 8.1-m-diameter Gemini North telescope at Mauna Kea, Hawaii, if only the diffraction effects of the telescope aperture limit the resolution? Assume 550 nm for the wavelength of light and 400,000 km for the distance to the moon.

Solution:

33 m

Exercise:

Problem:

A spy satellite is reputed to be able to resolve objects 10. cm apart while operating 197 km above the surface of Earth. What is the diameter of the aperture of the telescope if the resolution is only limited by the diffraction effects? Use 550 nm for light.

Exercise:

Problem:

Monochromatic light of wavelength 530 nm passes through a horizontal single slit of width $1.5\ \mu\text{m}$ in an opaque plate. A screen of dimensions $2.0\ \text{m}\ \times\ 2.0\ \text{m}$ is 1.2 m away from the slit. (a) Which way is the diffraction pattern spread out on the screen? (b) What are the angles of the minima with respect to the center? (c) What are the angles of the maxima? (d) How wide is the central bright fringe on the screen? (e) How wide is the next bright fringe on the screen?

Solution:

a. vertically; b. $\pm 20^\circ$, $\pm 44^\circ$; c. 0, $\pm 31^\circ$, $\pm 60^\circ$; d. 89 cm; e. 71 cm

Exercise:

Problem:

A monochromatic light of unknown wavelength is incident on a slit of width $20\ \mu\text{m}$. A diffraction pattern is seen at a screen 2.5 m away where the central maximum is spread over a distance of 10.0 cm. Find the wavelength.

Exercise:

Problem:

A source of light having two wavelengths 550 nm and 600 nm of equal intensity is incident on a slit of width $1.8\ \mu\text{m}$. Find the separation of the $m = 1$ bright spots of the two wavelengths on a screen 30.0 cm away.

Solution:

0.98 cm

Exercise:

Problem:

A single slit of width 2100 nm is illuminated normally by a wave of wavelength 632.8 nm. Find the phase difference between waves from the top and one third from the bottom of the slit to a point on a screen at a horizontal distance of 2.0 m and vertical distance of 10.0 cm from the center.

Exercise:

Problem:

A single slit of width $3.0\ \mu\text{m}$ is illuminated by a sodium yellow light of wavelength 589 nm. Find the intensity at a 15° angle to the axis in terms of the intensity of the central maximum.

Solution:

$I/I_0 = 0.041$

Exercise:

Problem:

A single slit of width 0.10 mm is illuminated by a mercury lamp of wavelength 576 nm. Find the intensity at a 10° angle to the axis in terms of the intensity of the central maximum.

Exercise:

Problem:

A diffraction grating produces a second maximum that is 89.7 cm from the central maximum on a screen 2.0 m away. If the grating has 600 lines per centimeter, what is the wavelength of the light that produces the diffraction pattern?

Solution:

340 nm

Exercise:

Problem:

A grating with 4000 lines per centimeter is used to diffract light that contains all wavelengths between 400 and 650 nm. How wide is the first-order spectrum on a screen 3.0 m from the grating?

Exercise:

Problem:

A diffraction grating with 2000 lines per centimeter is used to measure the wavelengths emitted by a hydrogen gas discharge tube. (a) At what angles will you find the maxima of the two first-order blue lines of wavelengths 410 and 434 nm? (b) The maxima of two other first-order lines are found at $\theta_1 = 0.097\,\mathrm{rad}$ and $\theta_2 = 0.132\,\mathrm{rad}$. What are the wavelengths of these lines?

Solution:

a. 0.082 rad and 0.087 rad; b. 480 nm and 660 nm

Exercise:

Problem:

For white light $(400\ \text{nm} < \lambda < 700\ \text{nm})$ falling normally on a diffraction grating, show that the second and third-order spectra overlap no matter what the grating constant *d* is.

Exercise:

Problem:

How many complete orders of the visible spectrum $(400\ \text{nm} < \lambda < 700\ \text{nm})$ can be produced with a diffraction grating that contains 5000 lines per centimeter?

Solution:

two orders

Exercise:

Problem:

Two lamps producing light of wavelength 589 nm are fixed 1.0 m apart on a wooden plank. What is the maximum distance an observer can be and still resolve the lamps as two separate sources of light, if the resolution is affected solely by the diffraction of light entering the eye? Assume light enters the eye through a pupil of diameter 4.5 mm.

Exercise:

Problem:

On a bright clear day, you are at the top of a mountain and looking at a city 12 km away. There are two tall towers 20.0 m apart in the city. Can your eye resolve the two towers if the diameter of the pupil is 4.0 mm? If not, what should be the minimum magnification power of the telescope needed to resolve the two towers? In your calculations use 550 nm for the wavelength of the light.

Solution:

yes and N/A

Exercise:

Problem:

Radio telescopes are telescopes used for the detection of radio emission from space. Because radio waves have much longer wavelengths than visible light, the diameter of a radio telescope must be very large to provide good resolution. For example, the radio telescope in Penticton, BC in Canada, has a diameter of 26 m and can be operated at frequencies as high as 6.6 GHz. (a) What is the wavelength corresponding to this frequency? (b) What is the angular separation of two radio sources that can be resolved by this telescope? (c) Compare the telescope’s resolution with the angular size of the moon.

(credit: modification of work by Jason Nishiyama)

Exercise:

Problem:

Calculate the wavelength of light that produces its first minimum at an angle of 36.9° when falling on a single slit of width $1.00\ \mu\text{m}$.

Solution:

600 nm

Exercise:

Problem:

(a) Find the angle of the third diffraction minimum for 633-nm light falling on a slit of width $20.0\ \mu\text{m}$. (b) What slit width would place this minimum at 85.0° ?

Exercise:

Problem:

As an example of diffraction by apertures of everyday dimensions, consider a doorway of width 1.0 m. (a) What is the angular position of the first minimum in the diffraction pattern of 600-nm light? (b) Repeat this calculation for a musical note of frequency 440 Hz (A above middle C). Take the speed of sound to be 343 m/s.

Solution:

a. $3.4\ \times\ 10^{-5\circ}$; b. 51°

Exercise:

Problem:

What are the angular positions of the first and second minima in a diffraction pattern produced by a slit of width 0.20 mm that is illuminated by 400 nm light? What is the angular width of the central peak?

Exercise:

Problem:

How far would you place a screen from the slit of the previous problem so that the second minimum is a distance of 2.5 mm from the center of the diffraction pattern?

Solution:

0.63 m

Exercise:

Problem:

How narrow is a slit that produces a diffraction pattern on a screen 1.8 m away whose central peak is 1.0 m wide? Assume $\lambda = 589\,\text{nm}$.

Exercise:

Problem:

Suppose that the central peak of a single-slit diffraction pattern is so wide that the first minima can be assumed to occur at angular positions of $\pm 90^\circ$. For this case, what is the ratio of the slit width to the wavelength of the light?

Solution:

1

Exercise:

Problem:

The central diffraction peak of the double-slit interference pattern contains exactly nine fringes. What is the ratio of the slit separation to the slit width?

Exercise:

Problem:

Determine the intensities of three interference peaks other than the central peak in the central maximum of the diffraction, if possible, when a light of wavelength 500 nm is incident normally on a double slit of width 1000 nm and separation 1500 nm. Use the intensity of the central spot to be $1\,\text{mW/cm}^2$.

Solution:

$0.17\ \text{mW/cm}^2$ for $m = 1$ only, no higher orders

Exercise:

Problem:

The yellow light from a sodium vapor lamp *seems* to be of pure wavelength, but it produces two first-order maxima at 36.093° and 36.129° when projected on a 10,000 line per centimeter diffraction grating. What are the two wavelengths to an accuracy of 0.1 nm?

Exercise:

Problem:

Structures on a bird feather act like a reflection grating having 8000 lines per centimeter. What is the angle of the first-order maximum for 600-nm light?

Solution:

28.7°

Exercise:

Problem:

If a diffraction grating produces a first-order maximum for the shortest wavelength of visible light at 30.0°, at what angle will the first-order maximum be for the largest wavelength of visible light?

Exercise:

Problem:

(a) What visible wavelength has its fourth-order maximum at an angle of 25.0° when projected on a 25,000-line per centimeter diffraction grating? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. 42.3 nm; b. This wavelength is not in the visible spectrum. c. The number of slits in this diffraction grating is too large. Etching in integrated circuits can be done to a resolution of 50 nm, so slit separations of 400 nm are at the limit of what we can do today. This line spacing is too small to produce diffraction of light.

Exercise:

Problem:

Consider a spectrometer based on a diffraction grating. Construct a problem in which you calculate the distance between two wavelengths of electromagnetic radiation in your spectrometer. Among the things to be considered are the wavelengths you wish to be able to distinguish, the number of lines per meter on the diffraction grating, and the distance from the grating to the screen or detector. Discuss the practicality of the device in terms of being able to discern between wavelengths of interest.

Exercise:

Problem:

An amateur astronomer wants to build a telescope with a diffraction limit that will allow him to see if there are people on the moons of Jupiter. (a) What diameter mirror is needed to be able to see 1.00-m detail on a Jovian moon at a distance of $7.50 \times 10^8\,\text{km}$ from Earth? The wavelength of light averages 600 nm. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

a. 549 km; b. This is an unreasonably large telescope. c. Unreasonable to assume diffraction limit for optical telescopes unless in space due to atmospheric effects.

Challenge Problems

Exercise:

Problem:

Blue light of wavelength 450 nm falls on a slit of width 0.25 mm. A converging lens of focal length 20 cm is placed behind the slit and focuses the diffraction pattern on a screen. (a) How far is the screen from the lens? (b) What is the distance between the first and the third minima of the diffraction pattern?

Exercise:

Problem:

(a) Assume that the maxima are halfway between the minima of a single-slit diffraction pattern. The use the diameter and circumference of the phasor diagram, as described in Intensity in Single-Slit Diffraction, to determine the intensities of the third and fourth maxima in terms of the intensity of the central maximum. (b) Do the same calculation, using [link].

Solution:

a. $\text{I} = 0.00500\,\text{I}_0,\ 0.00335\,\text{I}_0$; b. $\text{I} = 0.00500\,\text{I}_0,\ 0.00335\,\text{I}_0$

Exercise:

Problem:

(a) By differentiating [link], show that the higher-order maxima of the single-slit diffraction pattern occur at values of β that satisfy $\tan\beta = \beta$. (b) Plot $y = \tan\beta$ and $y = \beta$ versus β and find the intersections of these two curves. What information do they give you about the locations of the maxima? (c) Convince yourself that these points do not appear exactly at $\beta = \ n + \frac{1}{2}\ \pi$, where $n = 0,\ 1,\ 2,\ \ldots$, but are quite close to these values.

Exercise:

Problem:

What is the maximum number of lines per centimeter a diffraction grating can have and produce a complete first-order spectrum for visible light?

Solution:

12,800

Exercise:

Problem:

Show that a diffraction grating cannot produce a second-order maximum for a given wavelength of light unless the first-order maximum is at an angle less than 30.0°.

Exercise:

Problem:

A He-Ne laser beam is reflected from the surface of a CD onto a wall. The brightest spot is the reflected beam at an angle equal to the angle of incidence. However, fringes are also observed. If the wall is 1.50 m from the CD, and the first fringe is 0.600 m from the central maximum, what is the spacing of grooves on the CD?

Solution:

$1.58 \times 10^{-6}\,\text{m}$

Exercise:

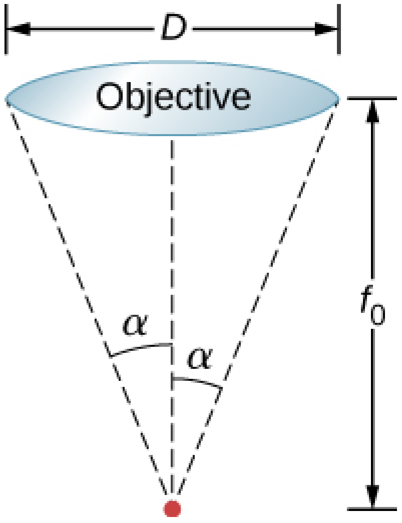
Problem:

Objects viewed through a microscope are placed very close to the focal point of the objective lens. Show that the minimum separation *x* of two objects resolvable through the microscope is given by

Equation:

$$x = \frac{1.22\lambda f_0}{D},$$

where f_0 is the focal length and D is the diameter of the objective lens as shown below.



Glossary

hologram
: three-dimensional image recorded on film by lasers; the word hologram means *entire picture* (from the Greek word *holo*, as in holistic)

holography
: process of producing holograms with the use of lasers

Units

Quantity	**Common Symbol**	**Unit**	**Unit in Terms of Base SI Units**
Acceleration	$\vec{\mathbf{a}}$	m/s^2	m/s^2
Amount of substance	n	**mole**	mol
Angle	θ, ϕ	radian (rad)	
Angular acceleration	$\vec{\alpha}$	rad/s^2	s^{-2}
Angular frequency	ω	rad/s	s^{-1}
Angular momentum	$\vec{\mathbf{L}}$	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}$	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}$
Angular velocity	$\vec{\omega}$	rad/s	s^{-1}
Area	A	m^2	m^2
Atomic number	Z		
Capacitance	C	farad (F)	$\mathrm{A}^2 \cdot \mathrm{s}^4/\mathrm{kg} \cdot \mathrm{m}^2$
Charge	q, Q, e	coulomb (C)	$\mathrm{A} \cdot \mathrm{s}$
Charge density:			
Line	λ	C/m	$\mathrm{A} \cdot \mathrm{s}/\mathrm{m}$
Surface	σ	C/m^2	$\mathrm{A} \cdot \mathrm{s}/\mathrm{m}^2$
Volume	ρ	C/m^3	$\mathrm{A} \cdot \mathrm{s}/\mathrm{m}^3$

Quantity	**Common Symbol**	**Unit**	**Unit in Terms of Base SI Units**
Conductivity	σ	$1/\Omega \cdot \mathrm{m}$	$\mathrm{A}^2 \cdot \mathrm{s}^3/\mathrm{kg} \cdot \mathrm{m}^3$
Current	I	**ampere**	A
Current density	$\vec{\mathbf{J}}$	$\mathrm{A/m}^2$	$\mathrm{A/m}^2$
Density	ρ	$\mathrm{kg/m}^3$	$\mathrm{kg/m}^3$
Dielectric constant	κ		
Electric dipole moment	$\vec{\mathbf{p}}$	$\mathrm{C} \cdot \mathrm{m}$	$\mathrm{A} \cdot \mathrm{s} \cdot \mathrm{m}$
Electric field	$\vec{\mathbf{E}}$	N/C	$\mathrm{kg} \cdot \mathrm{m}/\mathrm{A} \cdot \mathrm{s}^3$
Electric flux	Φ	$\mathrm{N} \cdot \mathrm{m}^2/\mathrm{C}$	$\mathrm{kg} \cdot \mathrm{m}^3/\mathrm{A} \cdot \mathrm{s}^3$
Electromotive force	ε	volt (V)	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{A} \cdot \mathrm{s}^3$
Energy	E,U,K	joule (J)	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}^2$
Entropy	S	J/K	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}^2 \cdot \mathrm{K}$
Force	$\vec{\mathbf{F}}$	newton (N)	$\mathrm{kg} \cdot \mathrm{m}/\mathrm{s}^2$
Frequency	f	hertz (Hz)	s^{-1}
Heat	Q	joule (J)	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}^2$
Inductance	L	henry (H)	$\mathrm{kg} \cdot \mathrm{m}^2/\mathrm{A}^2 \cdot \mathrm{s}^2$
Length:	ℓ, L	**meter**	m
Displacement	$\Delta x, \Delta\vec{\mathbf{r}}$		
Distance	d, h		
Position	$x, y, z, \vec{\mathbf{r}}$		

Quantity	Common Symbol	Unit	Unit in Terms of Base SI Units
Magnetic dipole moment	$\vec{\mu}$	$\text{N} \cdot \text{J/T}$	$\text{A} \cdot \text{m}^2$
Magnetic field	$\vec{\mathbf{B}}$	$\text{tesla (T)} = \quad \text{Wb/m}^2$	$\text{kg/A} \cdot \text{s}^2$
Magnetic flux	Φ_m	weber (Wb)	$\text{kg} \cdot \text{m}^2/\text{A} \cdot \text{s}^2$
Mass	*m*, *M*	**kilogram**	kg
Molar specific heat	*C*	$\text{J/mol} \cdot \text{K}$	$\text{kg} \cdot \text{m}^2/\text{s}^2 \cdot \text{mol} \cdot \text{K}$
Moment of inertia	*I*	$\text{kg} \cdot \text{m}^2$	$\text{kg} \cdot \text{m}^2$
Momentum	$\vec{\mathbf{p}}$	$\text{kg} \cdot \text{m/s}$	$\text{kg} \cdot \text{m/s}$
Period	*T*	s	s
Permeability of free space	μ_0	$\text{N/A}^2 = (\text{H/m})$	$\text{kg} \cdot \text{m/A}^2 \cdot \text{s}^2$
Permittivity of free space	ε_0	$\text{C}^2/\text{N} \cdot \text{m}^2 = (\text{F/m})$	$\text{A}^2 \cdot \text{s}^4/\text{kg} \cdot \text{m}^3$
Potential	*V*	$\text{volt (V)} = (\text{J/C})$	$\text{kg} \cdot \text{m}^2/\text{A} \cdot \text{s}^3$
Power	*P*	$\text{watt (W)} = (\text{J/s})$	$\text{kg} \cdot \text{m}^2/\text{s}^3$
Pressure	*p*	$\text{pascal (Pa)} = \quad \text{N/m}^2$	$\text{kg/m} \cdot \text{s}^2$
Resistance	*R*	$\text{ohm } (\Omega) = (\text{V/A})$	$\text{kg} \cdot \text{m}^2/\text{A}^2 \cdot \text{s}^3$
Specific heat	*c*	$\text{J/kg} \cdot \text{K}$	$\text{m}^2/\text{s}^2 \cdot \text{K}$
Speed	*v*	m/s	m/s
Temperature	*T*	**kelvin**	K
Time	*t*	**second**	s

Quantity	Common Symbol	Unit	Unit in Terms of Base SI Units
Torque	$\vec{\tau}$	$\mathrm{N \cdot m}$	$\mathrm{kg \cdot m^2/s^2}$
Velocity	$\vec{\mathbf{v}}$	m/s	m/s
Volume	V	$\mathrm{m^3}$	$\mathrm{m^3}$
Wavelength	λ	m	m
Work	W	$\text{joule}\,(\mathrm{J}) = (\mathrm{N \cdot m})$	$\mathrm{kg \cdot m^2/s^2}$

Units Used in Physics (Fundamental units in bold)

Conversion Factors

	m	**cm**	**km**
1 meter	1	10^2	10^{-3}
1 centimeter	10^{-2}	1	10^{-5}
1 kilometer	10^3	10^5	1
1 inch	2.540×10^{-2}	2.540	2.540×10^{-5}
1 foot	0.3048	30.48	3.048×10^{-4}
1 mile	1609	1.609×10^4	1.609
1 angstrom	10^{-10}		
1 fermi	10^{-15}		
1 light-year			9.460×10^{12}
	in.	**ft**	**mi**
1 meter	39.37	3.281	6.214×10^{-4}
1 centimeter	0.3937	3.281×10^{-2}	6.214×10^{-6}
1 kilometer	3.937×10^4	3.281×10^3	0.6214
1 inch	1	8.333×10^{-2}	1.578×10^{-5}
1 foot	12	1	1.894×10^{-4}
1 mile	6.336×10^4	5280	1

Length

Area

$1\ \text{cm}^2 = 0.155\ \text{in.}^2$

$1\ \text{m}^2 = 10^4\ \text{cm}^2 = 10.76\ \text{ft}^2$

$1\ \text{in.}^2 = 6.452\ \text{cm}^2$

$1\ \text{ft}^2 = 144\ \text{in.}^2 = 0.0929\ \text{m}^2$

Volume

$1\ \text{liter} = 1000\ \text{cm}^3 = 10^{-3}\ \text{m}^3 = 0.03531\ \text{ft}^3 = 61.02\ \text{in.}^3$

$1\ \text{ft}^3 = 0.02832\ \text{m}^3 = 28.32\ \text{liters} = 7.477\ \text{gallons}$

$1\ \text{gallon} = 3.788\ \text{liters}$

	s	**min**	**h**	**day**	**yr**
1 second	1	1.667×10^{-2}	2.778×10^{-4}	1.157×10^{-5}	3.169×10^{-8}
1 minute	60	1	1.667×10^{-2}	6.944×10^{-4}	1.901×10^{-6}
1 hour	3600	60	1	4.167×10^{-2}	1.141×10^{-4}
1 day	8.640×10^{4}	1440	24	1	2.738×10^{-3}
1 year	3.156×10^{7}	5.259×10^{5}	8.766×10^{3}	365.25	1

Time

	m/s	**cm/s**	**ft/s**	**mi/h**
1 meter/second	1	10^{2}	3.281	2.237
1 centimeter/second	10^{-2}	1	3.281×10^{-2}	2.237×10^{-2}
1 foot/second	0.3048	30.48	1	0.6818
1 mile/hour	0.4470	44.70	1.467	1

Speed

Acceleration

$1\ \text{m/s}^2 = 100\ \text{cm/s}^2 = 3.281\ \text{ft/s}^2$

$1\ \text{cm/s}^2 = 0.01\ \text{m/s}^2 = 0.03281\ \text{ft/s}^2$

$1\ \text{ft/s}^2 = 0.3048\ \text{m/s}^2 = 30.48\ \text{cm/s}^2$

$1\ \text{mi/h} \cdot \text{s} = 1.467\ \text{ft/s}^2$

	kg	**g**	**slug**	**u**
1 kilogram	1	10^3	6.852×10^{-2}	6.024×10^{26}
1 gram	10^{-3}	1	6.852×10^{-5}	6.024×10^{23}
1 slug	14.59	1.459×10^4	1	8.789×10^{27}
1 atomic mass unit	1.661×10^{-27}	1.661×10^{-24}	1.138×10^{-28}	1
1 metric ton	1000			

Mass

	N	**dyne**	**lb**
1 newton	1	10^5	0.2248
1 dyne	10^{-5}	1	2.248×10^{-6}
1 pound	4.448	4.448×10^5	1

Force

	Pa	**dyne/cm^2**	**atm**	**cmHg**	**lb/in.^{2}**
1 pascal	1	10	9.869×10^{-6}	7.501×10^{-4}	1.450×1
1 dyne/centimeter^2	10^{-1}	1	9.869×10^{-7}	7.501×10^{-5}	1.450×1
1 atmosphere	1.013×10^5	1.013×10^6	1	76	14.70
1 centimeter mercury*	1.333×10^3	1.333×10^4	1.316×10^{-2}	1	0.1934
1 pound/inch^2	6.895×10^3	6.895×10^4	6.805×10^{-2}	5.171	1
1 bar	10^5				
1 torr				1 (mmHg)	
*Where the acceleration due to gravity is $9.80665\ \mathrm{m/s^2}$ and the temperature is $0^\circ\mathrm{C}$					

Pressure

	J	**erg**	**ft.lb**
1 joule	1	10^7	0.7376
1 erg	10^{-7}	1	7.376×10^{-8}
1 foot-pound	1.356	1.356×10^7	1
1 electron-volt	1.602×10^{-19}	1.602×10^{-12}	1.182×10^{-19}
1 calorie	4.186	4.186×10^7	3.088
1 British thermal unit	1.055×10^3	1.055×10^{10}	7.779×10^2
1 kilowatt-hour	3.600×10^6		
	eV	**cal**	**Btu**
1 joule	6.242×10^{18}	0.2389	9.481×10^{-4}
1 erg	6.242×10^{11}	2.389×10^{-8}	9.481×10^{-11}
1 foot-pound	8.464×10^{18}	0.3239	1.285×10^{-3}
1 electron-volt	1	3.827×10^{-20}	1.519×10^{-22}
1 calorie	2.613×10^{19}	1	3.968×10^{-3}
1 British thermal unit	6.585×10^{21}	2.520×10^2	1

Work, Energy, Heat

Power

$1\ \mathrm{W} = 1\ \mathrm{J/s}$

$1\ \mathrm{hp} = 746\ \mathrm{W} = 550\ \mathrm{ft \cdot lb/s}$

$1\ \mathrm{Btu/h} = 0.293\ \mathrm{W}$

Angle

$1\ \mathrm{rad} = 57.30^\circ = 180^\circ/\pi$

$1^\circ = 0.01745\ \mathrm{rad} = \pi/180\ \mathrm{rad}$

$1\ \mathrm{revolution} = 360^\circ = 2\pi\ \mathrm{rad}$

$1\ \mathrm{rev/min\ (rpm)} = 0.1047\ \mathrm{rad/s}$

Fundamental Constants

Quantity	Symbol	Value
Atomic mass unit	u	$1.660\,538\,782\ (83)\ \times\ 10^{-27}\,\mathrm{kg}$ $931.494\,028\ (23)\,\mathrm{MeV}/c^2$
Avogadro’s number	N_A	$6.02214076\ \times\ 10^{23}$ reciprocal mole$\left(\mathrm{mol}^{-1}\right)$
Bohr magneton	$\mu_\mathrm{B}=\frac{e\hbar}{2m_e}$	$9.274\,009\,15\ (23)\ \times\ 10^{-24}\,\mathrm{J/T}$
Bohr radius	$a_0=\frac{\hbar^2}{m_e e^2 k_e}$	$5.291\,772\,085\,9\ (36)\ \times\ 10^{-11}\,\mathrm{m}$
Boltzmann’s constant	$k_\mathrm{B}=\frac{R}{N_\mathrm{A}}$	$1.380649\ \times\ 10^{-23}$ joule per kelvin$\left(\mathrm{J}\cdot\mathrm{K}^{-1}\right)$
Compton wavelength	$\lambda_\mathrm{C}=\frac{h}{m_e c}$	$2.426\,310\,217\,5\ (33)\ \times\ 10^{-12}\,\mathrm{m}$
Coulomb constant	$k_e=\frac{1}{4\pi\varepsilon_0}$	$8.987\,551\,788...\ \times\ 10^{9}\,\mathrm{N}\cdot\mathrm{m}^2/\mathrm{C}^2$ (exact)
Deuteron mass	m_d	$3.343\,583\,20\ (17)\ \times\ 10^{-27}\,\mathrm{kg}$ $2.013\,553\,212\,724\,(78)\,\mathrm{u}$ $1875.612\,859\,\mathrm{MeV}/c^2$
Electron mass	m_e	$9.109\,382\,15\ (45)\ \times\ 10^{-31}\,\mathrm{kg}$ $5.485\,799\,094\,3\,(23)\ \times\ 10^{-4}\,\mathrm{u}$ $0.510\,998\,910\ (13)\,\mathrm{MeV}/c^2$
Electron volt	eV	$1.602\,176\,487\ (40)\ \times\ 10^{-19}\,\mathrm{J}$

Quantity	**Symbol**	**Value**
Elementary charge	e	$1.602176634\ \times\ 10^{-19}\,\mathrm{C}$
Gas constant	R	$8.314\,472\ (15)\ \mathrm{J/mol\cdot K}$
Gravitational constant	G	$6.674\,28\ (67)\ \times\ 10^{-11}\,\mathrm{N\cdot m^2/kg^2}$
Neutron mass	m_n	$1.674\,927\,211\ (84)\ \times\ 10^{-27}\,\mathrm{kg}$ $1.008\,664\,915\,97\ (43)\ \mathrm{u}$ $939.565\,346\ (23)\ \mathrm{MeV}/c^2$
Nuclear magneton	$\mu_n = \frac{e\hbar}{2m_p}$	$5.050\,783\,24\ (13)\ \times\ 10^{-27}\,\mathrm{J/T}$
Permeability of free space	μ_0	$4\pi\ \times\ 10^{-7}\,\mathrm{T\cdot m/A}$ (exact)
Permittivity of free space	$\varepsilon_0 = \frac{1}{\mu_0 c^2}$	$8.854\,187\,817...\ \times\ 10^{-12}\,\mathrm{C^2/N\cdot m^2}$ (exact)
Planck's constant	h $\hbar = \frac{h}{2\pi}$	$6.62607015\ \times\ 10^{-34}\,\mathrm{kg\cdot m^2\cdot s^{-1}}$ $1.05457182\ \times\ 10^{-34}\,\mathrm{kg\cdot m^2\cdot s^{-1}}$
Proton mass	m_p	$1.672\,621\,637\ (83)\ \times\ 10^{-27}\,\mathrm{kg}$ $1.007\,276\,466\,77\ (10)\ \mathrm{u}$ $938.272\,013\ (23)\ \mathrm{MeV}/c^2$
Rydberg constant	R_H	$1.097\,373\,156\,852\,7\ (73)\ \times\ 10^{7}\,\mathrm{m^{-1}}$
Speed of light in vacuum	c	$2.997\,924\,58\ \times\ 10^{8}\,\mathrm{m/s}$ (exact)

Fundamental Constants*Note:* These constants are the values recommended in 2006 by CODATA, based on a least-squares adjustment of data from different measurements. The numbers in parentheses for the values represent the uncertainties of the last two digits.

Useful combinations of constants for calculations:

$hc = 12{,}400\,\text{eV} \cdot \text{Å} = 1240\,\text{eV} \cdot \text{nm} = 1240\,\text{MeV} \cdot \text{fm}$

$\hbar c = 1973\,\text{eV} \cdot \text{Å} = 197.3\,\text{eV} \cdot \text{nm} = 197.3\,\text{MeV} \cdot \text{fm}$

$k_e e^2 = 14.40\,\text{eV} \cdot \text{Å} = 1.440\,\text{eV} \cdot \text{nm} = 1.440\,\text{MeV} \cdot \text{fm}$

$k_\text{B} T = 0.02585\,\text{eV at } T = 300\,\text{K}$

Astronomical Data

Celestial Object	**Mean Distance from Sun (million km)**	**Period of Revolution (d = days) (y = years)**	**Period of Rotation at Equator**	**Eccentricity of Orbit**
Sun	−	−	27 d	−
Mercury	57.9	88 d	59 d	0.206
Venus	108.2	224.7 d	243 d	0.007
Earth	149.6	365.26 d	23 h 56 min 4 s	0.017
Mars	227.9	687 d	24 h 37 min 23 s	0.093
Jupiter	778.4	11.9 y	9 h 50 min 30 s	0.048
Saturn	1426.7	29.5 6	10 h 14 min	0.054
Uranus	2871.0	84.0 y	17 h 14 min	0.047
Neptune	4498.3	164.8 y	16 h	0.009

Celestial Object	**Mean Distance from Sun (million km)**	**Period of Revolution (d = days) (y = years)**	**Period of Rotation at Equator**	**Eccentricity of Orbit**
Earth’s Moon	149.6 (0.386 from Earth)	27.3 d	27.3 d	0.055
Celestial Object	**Equatorial Diameter (km)**	**Mass (Earth = 1)**	**Density (g/cm^3)**	
Sun	1,392,000	333,000.00	1.4	
Mercury	4879	0.06	5.4	
Venus	12,104	0.82	5.2	
Earth	12,756	1.00	5.5	
Mars	6794	0.11	3.9	
Jupiter	142,984	317.83	1.3	
Saturn	120,536	95.16	0.7	
Uranus	51,118	14.54	1.3	
Neptune	49,528	17.15	1.6	
Earth’s Moon	3476	0.01	3.3	

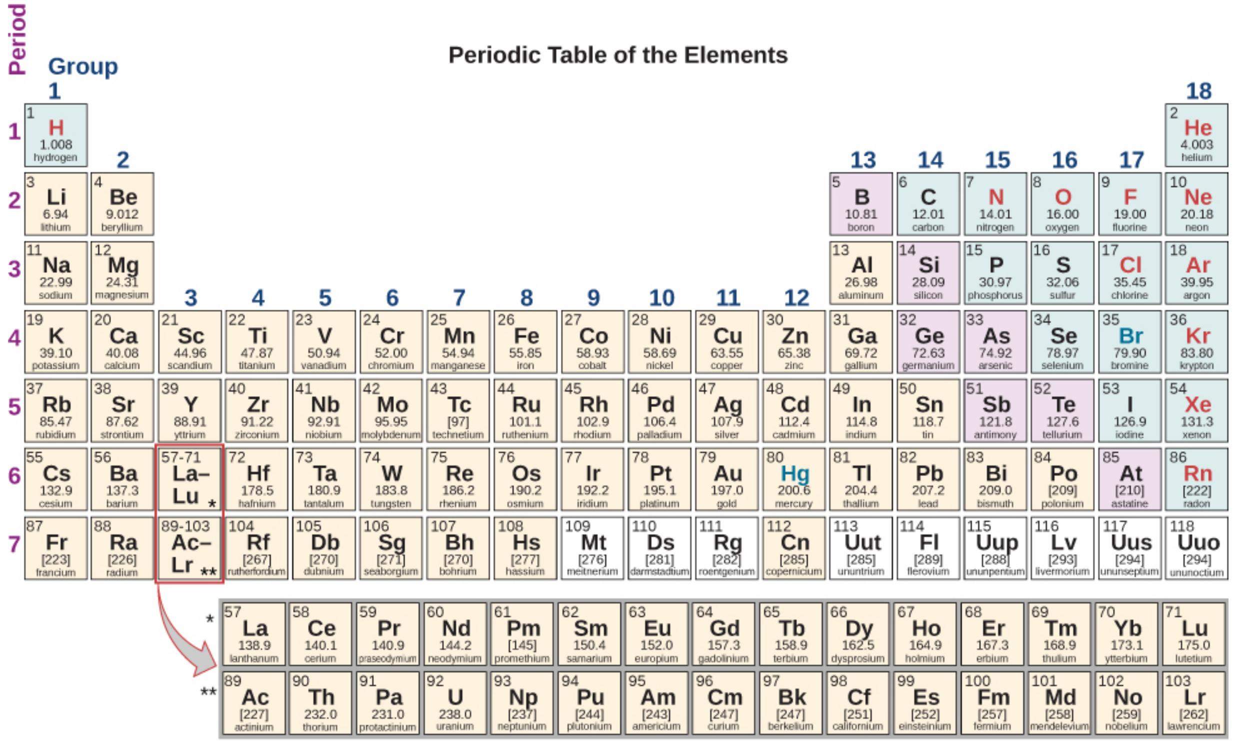
Astronomical Data

Other Data:

Mass of Earth: 5.97×10^{24} kg

Mass of the Moon: 7.36×10^{22} kg

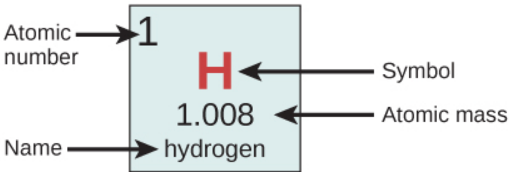
Mass of the Sun: 1.99×10^{30} kg



Periodic Table of the Elements

Period \ Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H 1.008 hydrogen																	2 He 4.003 helium
2	3 Li 6.94 lithium	4 Be 9.012 beryllium											5 B 10.81 boron	6 C 12.01 carbon	7 N 14.01 nitrogen	8 O 16.00 oxygen	9 F 19.00 fluorine	10 Ne 20.18 neon
3	11 Na 22.99 sodium	12 Mg 24.31 magnesium											13 Al 26.98 aluminum	14 Si 28.09 silicon	15 P 30.97 phosphorus	16 S 32.06 sulfur	17 Cl 35.45 chlorine	18 Ar 39.95 argon
4	19 K 39.10 potassium	20 Ca 40.08 calcium	21 Sc 44.96 scandium	22 Ti 47.87 titanium	23 V 50.94 vanadium	24 Cr 52.00 chromium	25 Mn 54.94 manganese	26 Fe 55.85 iron	27 Co 58.93 cobalt	28 Ni 58.69 nickel	29 Cu 63.55 copper	30 Zn 65.38 zinc	31 Ga 69.72 gallium	32 Ge 72.63 germanium	33 As 74.92 arsenic	34 Se 78.97 selenium	35 Br 79.90 bromine	36 Kr 83.80 krypton
5	37 Rb 85.47 rubidium	38 Sr 87.62 strontium	39 Y 88.91 yttrium	40 Zr 91.22 zirconium	41 Nb 92.91 niobium	42 Mo 95.95 molybdenum	43 Tc [97] technetium	44 Ru 101.1 ruthenium	45 Rh 102.9 rhodium	46 Pd 106.4 palladium	47 Ag 107.9 silver	48 Cd 112.4 cadmium	49 In 114.8 indium	50 Sn 118.7 tin	51 Sb 121.8 antimony	52 Te 127.6 tellurium	53 I 126.9 iodine	54 Xe 131.3 xenon
6	55 Cs 132.9 cesium	56 Ba 137.3 barium	57-71 La– Lu *	72 Hf 178.5 hafnium	73 Ta 180.9 tantalum	74 W 183.8 tungsten	75 Re 186.2 rhenium	76 Os 190.2 osmium	77 Ir 192.2 iridium	78 Pt 195.1 platinum	79 Au 197.0 gold	80 Hg 200.6 mercury	81 Tl 204.4 thallium	82 Pb 207.2 lead	83 Bi 209.0 bismuth	84 Po [209] polonium	85 At [210] astatine	86 Rn [222] radon
7	87 Fr [223] francium	88 Ra [226] radium	89-103 Ac– Lr **	104 Rf [267] rutherfordium	105 Db [270] dubnium	106 Sg [271] seaborgium	107 Bh [270] bohrium	108 Hs [277] hassium	109 Mt [276] meitnerium	110 Ds [281] darmstadtium	111 Rg [282] roentgenium	112 Cn [285] copernicium	113 Uut [285] ununtrium	114 Fl [289] flerovium	115 Uup [288] ununpentium	116 Lv [293] livermorium	117 Uus [294] ununseptium	118 Uuo [294] ununoctium

*	57 La 138.9 lanthanum	58 Ce 140.1 cerium	59 Pr 140.9 praseodymium	60 Nd 144.2 neodymium	61 Pm [145] promethium	62 Sm 150.4 samarium	63 Eu 152.0 europium	64 Gd 157.3 gadolinium	65 Tb 158.9 terbium	66 Dy 162.5 dysprosium	67 Ho 164.9 holmium	68 Er 167.3 erbium	69 Tm 168.9 thulium	70 Yb 173.1 ytterbium	71 Lu 175.0 lutetium
**	89 Ac [227] actinium	90 Th 232.0 thorium	91 Pa 231.0 protactinium	92 U 238.0 uranium	93 Np [237] neptunium	94 Pu [244] plutonium	95 Am [243] americium	96 Cm [247] curium	97 Bk [247] berkelium	98 Cf [251] californium	99 Es [252] einsteinium	100 Fm [257] fermium	101 Md [258] mendelevium	102 No [259] nobelium	103 Lr [262] lawrencium



Color Code

🟨 Metal	**Solid**
🟪 Metalloid	**Liquid**
🟦 Nonmetal	**Gas**

The Greek Alphabet

Name	**Capital**	**Lowercase**	**Name**	**Capital**	**Lowercase**
Alpha	A	α	Nu	N	ν
Beta	B	β	Xi	Ξ	ξ
Gamma	Γ	γ	Omicron	O	o
Delta	Δ	δ	Pi	Π	π
Epsilon	E	ε	Rho	P	ρ
Zeta	Z	ζ	Sigma	Σ	σ
Eta	H	η	Tau	T	τ
Theta	Θ	θ	Upsilon	ϒ	υ
Iota	I	ι	Phi	Φ	ϕ
Kappa	K	κ	Chi	X	χ
Lambda	Λ	λ	Psi	ψ	ψ
Mu	M	μ	Omega	Ω	ω

The Greek Alphabet